(12)特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2005 年9 月29 日 (29.09.2005)

PCT

(10) 国際公開番号
WO 2005/091170 A1

(51) 国際特許分類[7]: G06F 17/30, 17/27, 19/00

(21) 国際出願番号: PCT/JP2005/005440

(22) 国際出願日: 2005 年3 月17 日 (17.03.2005)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2004-079077 2004 年3 月18 日 (18.03.2004) JP

(71) 出願人 (米国を除く全ての指定国について): 日本電気株式会社 (NEC CORPORATION) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

(72) 発明者; および
(75) 発明者/出願人 (米国についてのみ): 坂尾 要祐 (SAKAO, Yousuke) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP). 佐藤 研治 (SATOH, Kenji) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP). 赤峯 享 (AKAMINE, Susumu) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP).

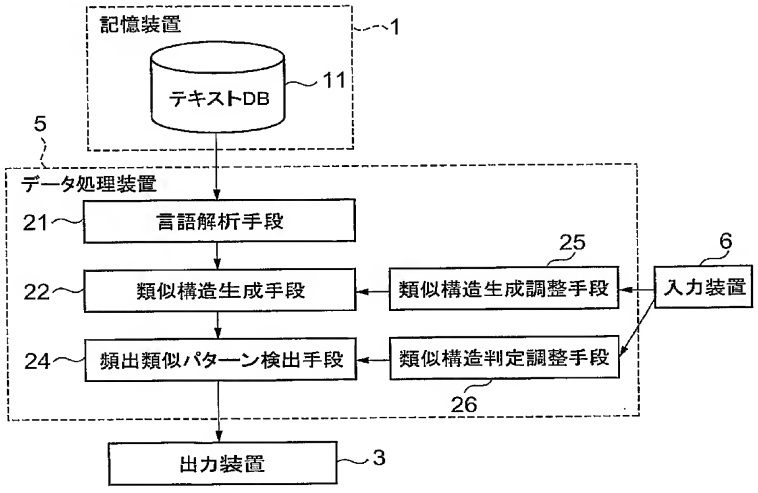
(74) 代理人: 池田 憲保, 外(IKEDA, Noriyasu et al.); 〒1050003 東京都港区西新橋一丁目４番１０号 第３森ビル Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM,

[続葉有]

(54) Title: TEXT MINING DEVICE, METHOD THEREOF, AND PROGRAM

(54) 発明の名称: テキストマイニング装置、その方法及びプログラム



1 STORAGE DEVICE
11 TEXT DB
5 DATA PROCESSING DEVICE
21 LANGUAGE ANALYSIS MEANS
22 SIMILAR STRUCTURE GENERATION MEANS
24 FREQUENCY SIMILAR PATTER DETECTION MEANS
25 SIMILAR STRUCTURE GENERATION ADJUSTMENT MEANS
26 SIMILAR STRUCTURE JUDGMENT ADJUSTMENT MEANS
6 INPUT DEVICE
3 OUTPUT DEVICE

(57) Abstract: Language analysis means (21) analyzes each of the texts read from a text DB (11) and generates a sentence structure as an analysis result. Similar structure generation adjustment means (25) generates a specification item for specifying whether the input from an input device has the same structure for each of the types of the sentence structure. Similar structure judgment adjustment means (26) generates a specification item whether to ignore the difference in value for each of the types of the attribute values according to the input from an input device (6). According to the specification item from the similar structure generation adjustment means (25), similar structure generation means (22) generates a similar structure of the partial structure constituting the sentence structure obtained by the language analysis means (21) and sets each of the similar structures generated as an equivalent value of the partial structure of each of the elements generated. According to the specification item given from the similar structure judgment adjustment means (26), frequent similar pattern detection means (24) ignores the attribute value, detects a frequent pattern from a set of the equivalent values from the similar structure generation means (22), and outputs it to an output device (3).

(57) 要約: 言語解析手段２１はテキストＤＢ１１から読み込んだ各テキストの解析を行い解析結果として文構造を生成する。類似構造生成調整手段２５は入力装置からの入力より文

[続葉有]

WO 2005/091170 A1

DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告書
— 補正書

２文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

構造の差異の種別毎に同一構造と判定するか否かを指定する指定項目を生成する。類似構造判定調整手段２６は入力装置６からの入力より属性値の種別毎に値の差異を無視するか否かを指定する指定項目を生成する。類似構造生成手段２２は類似構造生成調整手段２５からの指定項目に従い言語解析手段２１が得た文構造を構成する部分構造の類似構造を生成し生成した各類似構造を夫々の生成元の部分構造の同値類とする。頻出類似パターン検出手段２４は類似構造判定調整手段２６より与えられた指定項目に従い属性値を無視し類似構造生成手段２２からの同値類の集合より頻出パターンを検出し出力装置３に出力する。

明　細　書

テキストマイニング装置、その方法及びプログラム

技術分野

本発明は、構文解析などを用いて、コンピュータ上に蓄積される電子化テキストを構造化して分析を行うテキストマイニング装置、テキストマイニング方法及びテキストマイニング用プログラムに関し、特に、意味の類似した文の構造を同一の構造と判定して分析を行うことができるテキストマイニング装置、テキストマイニング方法及びテキストマイニング用プログラムに関する。

背景技術

テキストマイニング装置の一例として、図１に示すような構成が知られている（特許文献：特開２００１－８４２５０号公報（第４、５頁、第３図）参照）。図１に示すように、この従来のテキストマイニング装置は、基本辞書記憶部と、文書データ記憶部と、分野依存辞書記憶部と、言語特徴分析装置と、言語解析装置と、パターン抽出装置と、頻出パターン表示装置とを備えている。

図１に示した従来のテキストマイニング装置は、概略、つぎのように動作する。まず、言語特徴分析装置によって基本辞書と文書データとから分野依存辞書を作成し、言語解析装置によって基本辞書と分野依存辞書と文書データから構文木等の構造を作成する。パターン抽出装置は、この構造を用いて頻出パターンを抽出し、この頻出パターンに合致する文書データ中の文書を、頻出パターン適合文書記憶部に記憶させると同時に、この頻出パターンを出力する。

一般的に、言語解析装置によって作成される構造として、例えば、

（Ａ１）文中の文節を、構造の節点で表し、

（Ａ２）付属語情報を、節点の属性値で表し、

（Ａ３）係り受け関係を、係り元の節点から係り先の節点への有向枝で表し、

（Ａ４）表層格の情報を、有向枝の属性値で表す

という構造が良く用いられる。

ここで、付属語情報とは、進行や完了などの時制、容易や困難などのモダリティ、及び否定などの付属的な概念である。前記付属語情報は付属語によって文節に付加される情報をいう。

図２に、この形式で表された「彼は車種Ａが価格を下げたのを知らない」という文の構文構造の一例を示す。文の文節、「彼」、「車種Ａ」、「価格」、「下げる」、「知る」は節点で表わされる。付属語情報は節点の属性値で表される（節点「知る」の属性値として、付属語情報：否定)。係り受け関係は、係り元の節点から係り先への有向枝で表わされる（例えば「彼」→「知る」)。表層格の情報は有向枝の属性値で表される（例えば「彼」→「知る」の有向枝の属性値として「表層格：は」)。

また、構造中のこれらの情報は、全て属性値を持たないラベル付きの節点と、属性値を持たない有向枝のみからなる構造で表現することも可能である。図３に、この形式で表された「彼は車種Ａが価格を下げたのを知らない」という文の構文構造の例を示す。

文の文節「彼」、「車種Ａ」、「価格」、「下げる」、「知る」は、属性値を持たないラベル付きの節点で表わされ（例えば節点「彼」には「表層格：は」のラベルが付加され、「下げる」には、ラベル「付属語情報：完了」、「表層格：を」が付加されている)、係り元の節点から係り先への有向枝は属性値を持たない有向枝とされる。

上記した従来のシステムは下記記載の問題点を有している。なお、以下の問題点及びその解析は、本願発明者らによる研究・検討結果に基づくものである。図４Ａ～４Ｄ、図５Ａ及び図５Ｂの内容は、問題の在り処を具体的に説明するために、本願発明者らが提示したものである。

第１の問題点として、頻出パターン検出の際に、意味が類似し、かつ連結構造が異なっている構造は、全く別のパターンとして判定されてしまうということが挙げられる。

連結構造とは、構造の節点と単語文字列及び有向枝の連結関係と、向きにのみ注目し、付属的な属性情報を省略した構造のことをいう。

上記第１の問題点が生じる理由は、従来のテキストマイニング装置は、連結構造が異なり、類似した意味を持つ構造を同一と判定する手段を具備していないためである。

属性値を用いた文構造を用いる際に、連結構造が異なり、類似した意味を持つ構造の差異の例として、

（Ｂ１）係り受けの向きの差異、

（Ｂ２）係り受けの順序の差異、

（Ｂ３）同義語の置換による差異、及び、

（Ｂ４）並列の構文構造と意味構造の差異

などが挙げられる。

図４Ａ～４Ｄに、これらの連結構造による構造の差異の例を示す。属性値を用いない文構造を用いる際、あらゆる意味の類似した構造の差異は、連結構造の差異で表現される。

図４Ａに示す例では、意味の類似した「速いのは車種Ａ」と「車種Ａは速い」の連結構造において、係り元と係り先が相違している。

図４Ｂに示す例では、意味の類似した「速く安い車種Ａ」と「安く速い車種Ａ」の連結構造において、係り元の「速い」と「安い」の節点の順序関係が、相違している。

図４Ｃに示す例では、意味の類似した「車種Ａは速い」と「車種Ａは高速だ」のそれぞれの連結構造において、係り先の「速い」と「高速」が相違している。

図４Ｄに示す例では、「車種Ａと車種Ｂは速い」の構文構造と意味構造を表わしている。図４Ｄにおいて、係り元「車種Ａ」が「車種Ｂ」に係り「車種Ｂ」が「速い」に係る連結構造と、係り元「車種Ａ」と「車種Ｂ」から係り先「速い」への有向枝を有する連結構造がある。

第２の問題点として、頻出パターン検出の際に異なる属性値を持ち、かつ類似した意味を持つ構造は、全く別のパターンとして判定されてしまうということが挙げられる。

その理由は、従来のテキストマイニング装置では、異なる属性値を持つ構造を、同一と判定することについて、何ら考慮されていないためである。

属性値を用いた文構造を用いる際に、属性値が異なり、かつ類似した意味を持つ構造の差異の例として、付属語情報の差異、表層格の差異などが挙げられる。図５Ａ及び図５Ｂに、これらの属性値による構造の差異の例を示す。

図５Ａに示す例では、類似した意味を持つ「車種Ａは加速」と「車種Ａの加速」の連結構造において、有向枝の表層格が相違している。

図５Ｂに示す例では、類似した意味を持つ「車種Ａは速い」と「車種Ａは速かった」の連結構造において、係り先の節点「速い」の付属語情報が相違している。

第３の問題点として、テキストマイニング装置の使用者（ユーザ）がどこまで類似した構造を同一な構造と判定して頻出パターンの検出を行うのかを調整できないことが挙げられる。

その理由は、従来のテキストマイニング装置では、使用者が頻出パターン検出の際にどのような構造を同一と判定するかを調整することについて、何ら考慮されていないためである。

したがって、本発明の目的は、類似した意味を持ち、かつ連結構造の異なる構造を、同一のパターンと判定して頻出パターン等の検出を行うテキストマイニング装置及び方法並びにプログラムを提供することにある。

本発明の他の目的は、類似した意味を持ち属性値の異なる構造を同一な構造と判定して頻出パターン検出を行うかを調整できるテキストマイニング装置及び方法並びにプログラムを提供することにある。

本発明のさらに他の目的は、テキストマイニングの使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整できるテキストマイニング装置及び方法並びにプログラムを提供することにある。

発明の開示

本願で開示される発明は、上記目的を達成するため、概略以下の構成とされる。

本発明の第１の態様に係るテキストマイニング装置は、入力した文書から文構造を作成する手段と、前記文構造の部分構造に対して予め定められた所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する手段と、前記意味の類似したパターンを同一パターンと判定してパターン検出を行う手段と、を備えている。

本発明において、前記類似構造を生成する手段は、前記文構造について並列変形を行う手段と、前記文構造の部分構造を生成する手段と、前記文書構造及び／又は

部分構造の有向枝の無向枝化を行う手段と、同義語辞書を参照して前記文書構造及び／又は部分構造中の同義語の置換を行う手段と、前記文書構造及び／又は部分構造における順序木の無順序木化を行う手段と、を備え、前記類似構造を前記部分構造の同値類とする。同値類とは、構造の集合でその各要素を同一の構造として扱うものをいい、二つの同値類に一つでも、同一の要素が含まれる時には、その二つの同値類を同一の同値類と判定する。本発明によれば、生成された類似構造を生成元の文構造の同値類として扱い、頻出パターン検出を行う。

本発明の第２の態様に係るテキストマイニング装置は、第１の態様に係るテキストマイニング装置の構成に含まれる頻出パターン検出手段に代わり、構造中の属性値の差異を無視して、頻出パターンの検出を行う頻出類似パターン検出手段を備え、属性値の異なる類似した構造を同一な構造と判定して頻出パターンの検出を行う。本発明によれば、構造中の属性値が異なる類似した構造を同一と判定して頻出パターン検出を行う。

本発明の第３の態様に係るテキストマイニング装置は、テキストマイニングの対象となる文書の集まりを記憶する記憶部と、前記記憶部の前記文書を解析して文構造を取得する解析部と、使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目を生成する類似構造生成調整部と、使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成し類似構造判定調整部と、前記類似構造生成調整部によって生成された第１の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する類似構造生成部と、前記類似構造生成部によって生成された類似構造を生成元の部分構造の同値類として扱い、前記類似構造判定調整部の第２の指定項目に従い、属性値の差異を無視しながら、頻出パターンの検出を行う類似パターン検出部と、を備えている。本発明によれば、構造の同一性の判定を調整するための指定の入力を受け付ける。

本発明のさらに他の態様に係る方法は、

入力した文書から文構造を作成する工程と、

前記文構造の部分構造に対する所定の変換操作を行うことで、前記部分構造と意

味の類似したパターンの類似構造を作成する工程と、

前記意味の類似したパターンを同一パターンと判定してパターン検出を行う工程とを含む。

本発明のさらに他の態様に係る方法は、テキストマイニングの対象となるテキストの集まりを記憶する記憶部のテキストを解析して文構造を取得する工程と、

前記文構造の部分構造に対して意味的に類似しパターンの類似構造を生成する工程と、

生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異を無視しながらパターンの検出を行う工程とを含む。

本発明のさらに他の態様に係る方法は、テキストマイニングの対象となるテキストの集まりを記憶する記憶部のテキストを解析して文構造を取得する工程と、

入力装置から入力された使用者の入力情報から、文構造（連結構造）の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目と、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成するステップと、

文構造（連結構造）の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目に従い、前記文構造の部分構造に対して意味的に類似した構造を生成する工程と、

生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目に従い、属性値の差異を無視しながら頻出パターンの検出を行う工程と、を含む。

本発明のさらに他の態様に係るプログラムは、テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となるテキストの集まりを記憶する記憶部の前記テキストを解析して文構造を取得する処理と、

前記処理で解析して得られた文構造の部分構造に対して、意味的に類似した構造を生成する処理と、

生成された類似構造を、生成元の部分構造の同値類として扱い、頻出パターンの検出を行う処理と、

を実行させるプログラムよりなる。

図面の簡単な説明

図１は従来の技術の構成を示す図である。

図２は属性値を用いる形式で表された「彼は私が本を買ったのを知らない」という文の構文構造の例を示す図である。

図３は属性値を用いない形式で表された「彼は私が本を買ったのを知らない」という文の構文構造の例を示す図である

図４Ａは連結構造が異なり類似した意味を持つ構造の差異の例を示す図であり、係り受けの向きの差異を示した図である。

図４Ｂは連結構造が異なり類似した意味を持つ構造の差異の例を示す図であり、係り受けの順序の差異を示した図である。

図４Ｃは連結構造が異なり類似した意味を持つ構造の差異の例を示す図であり、同義語の置換による差異を示す図である。

図４Ｄは連結構造が異なり類似した意味を持つ構造の差異の例を示す図であり、並列の構文構造と意味構造の差異を示す図である。

図５Ａは属性値が異なり類似した意味を持つ構造の差異の複数の例を示す図であり、付属語情報の差異を示す図である。

図５Ｂは属性値が異なり類似した意味を持つ構造の差異の複数の例を示す図であり、表層格の差異を示す図である。

図６は本発明の第１の実施の形態の構成を示す図である。

図７は第１の実施の形態の動作を説明するための流れ図である。

図８は本発明の実施の形態における類似構造生成手段２２の動作を説明するための流れ図である。

図９は本発明の第２の実施の形態の構成を示す図である。

図１０は本発明の第２の実施の形態の動作を説明するための流れ図である。

図１１は本発明の第３の実施の形態の構成を示す図である。

図１２は本発明の第３の実施の形態の動作を説明するための流れ図である。

図１３は本発明の第３の実施の形態における類似構造生成手段２２の動作を説

明するための流れ図である。

図１４は本発明の第４の実施の形態の構成を示す図である。

図１５は本発明の第１～第３実施例で使用するテキストＤＢ中のテキスト集合の例を示す図である。

図１６Ａは言語解析手段２１で得られる文１の文構造を示す図である。

図１６Ｂは言語解析手段２１で得られる文２の文構造を示す図である。

図１６Ｃは言語解析手段２１で得られる文３の文構造を示す図である。

図１７は本発明の第１～第３の実施例において使用する、同義語辞書の構造を示す図である。

図１８は本発明の第１～第３の実施例において、図８のステップＡ２－１における処理を示す図である。

図１９は本発明の第１～第３の実施例において、図８のステップＡ２－２における処理を示す図である。

図２０Ａは部分構造２ａ－０に対する無効枝化処理（ステップＡ２－３）を示す図である。

図２０Ｂは部分構造２ｃ－０に対する無効枝化処理（ステップＡ２－３）を示す図である。

図２０Ｃは部分構造２ａ－１に対する無効枝化処理（ステップＡ２－３）を示す図である。

図２０Ｄは部分構造２ｇ－０に対する無効枝化処理（ステップＡ２－３）を示す図である。

図２０Ｅは部分構造２ｂ－０に対する無効枝化処理（ステップＡ２－３）を示す図である。

図２１は本発明の第１～第３の実施例において、図８のステップＡ２－６における処理を示す図である。

図２２は本発明の第１、第２の実施例において、類似構造生成手段２２が文３の文構造の全体からなる部分構造３ａ－０の類似構造を生成する処理を示す図である。

図２３は本発明の第１～第３の実施例において文１の文構造から生成される部

分構造の同値類を示す図である。

図２４は本発明の第１～第３の実施例において文２の文構造から生成される部分構造の同値類を示す図である。

図２５は本発明の第１、第２の実施例において、文３の文構造から生成される部分構造の同値類を示す図である。

図２６は本発明の第１の実施例において、図２３～２５に示す同値類の集合から検出される頻出パターンを示す図である。

図２７は本発明の第２の実施例において、図２３～２５に示す同値類の集合から検出される頻出パターンを示す図である。

図２８は本発明の第３の実施例において、類似構造生成手段２２が文３の文構造の全体からなる部分構造３ａ－０の類似構造を生成する処理を示す図である。

図２９は本発明の第３の実施例において、文３の文構造から生成される部分構造の同値類を示す図である。

図３０は本発明の第３の実施例において、図２３、２４及び図２９に示す同値類の集合から検出される頻出パターンを示す図である。

発明を実施するための最良の形態

以下、発明を実施するための最良の形態について図面を参照して詳細に説明する。

図６を参照すると、本発明の第１の実施の形態に係る装置は、情報を記憶する記憶装置１と、プログラム制御により動作するデータ処理装置２と、検出されたパターンを出力する出力装置３と、を有している。記憶装置１はテキストデータベース（ＤＢ）１１を含む。テキストＤＢ１１は、テキストマイニングの対象となるテキストの集合を記憶している。

データ処理装置２は、言語解析手段２１と、類似構造生成手段２２と、頻出パターン検出手段２３を含む。これらの手段はそれぞれ、おおむね以下のように動作する。

言語解析手段２１は、テキストＤＢ１１からテキスト集合を読み込み、その結果、集合中の各テキストを解析して文構造を得る。

類似構造生成手段２２は、言語解析手段２１から送出された文構造の集合中の各

文構造を構成する全ての部分構造を抽出し、前記各部分構造の全ての類似構造を生成して、その結果、類似構造と生成元の部分構造を同値類とする。

頻出パターン検出手段２３は、類似構造生成手段２２から送出された部分構造の同値類の集合から頻出するパターンを検出し、出力装置３へ送出する。

図７は、本実施形態の動作を説明するための流れ図である。次に、図６及び図７を参照して、本発明の第１の実施形態に係る装置の動作について詳細に説明する。

まず、言語解析手段２１が、テキストＤＢ１１から、テキスト集合を読み込む。言語解析手段２１は、テキスト集合中の各テキストに対し解析を行い、解析結果として、文構造を生成し、類似構造生成手段２２に送出する（図７のステップＡ１）。

次に、類似構造生成手段２２は、与えられた文構造の集合中の部分構造の全ての類似構造を生成し、その結果、類似構造を生成元の部分構造の同値類とする。類似構造生成手段２２は、その後、同値類の集合を頻出パターン検出手段２３に送出する（図７のステップＡ２）。

さらに、頻出パターン検出手段２３は、与えられた部分構造の同値類から、頻出パターンの検出を行う（図７のステップＡ３）。

頻出パターン検出手段２３は、検出した頻出パターンを出力装置３に出力する（図７のステップＡ４）。

図８は、図７のステップＡ２における、類似構造生成手段２２の動作の詳細なフローチャートを示す図である。

図８を参照すると、類似構造生成手段２２は、まず並列構文の構文構造と意味構造の違いに対応するための「並列の変形」を行う（図８のステップＡ２－１）。

次に、文構造全体だけではなく部分構造からもパターン検出を行うための「部分構造の生成」を行う（図８のステップＡ２－２）。

次に、係り受けの向きの差異に対応するための「有向枝の無向枝化」を行う（図８のステップＡ２－３）。

次に、同義語の差異に対応するための「同義語の置換」を行う（図８のステップＡ２－４）。

その係り受けの順序の違いに対応するための「順序木の無順序木化」を行う（図８のステップＡ２－５）。

最後に、類似構造を、生成元の部分構造の同値類の要素とすることで、「同値類の生成」を行う（図８のステップＡ２－６）。

以下、本発明の第１の実施の形態に係る装置の作用効果について説明する。

本実施の形態に係る装置は、類似構造生成手段２２が生成した類似構造を、元の構造の同値類として扱い、頻出パターン検出を行うように構成されている。このため、連結構造は異なるが、類似した意味を持つ構造を、同一の構造と判定して、頻出パターンを検出できる。

次に、本発明の第２の実施の形態について図面を参照して詳細に説明する。

図９を参照すると、本発明の第２の実施の形態に係る装置は、第１の実施の形態に係る装置と、データ処理装置４が、データ処理装置２の頻出パターン検出手段２３の代わりに頻出類似パターン検出手段２４を備えている以外は同じである。言語解析手段２１、類似構造生成手段２２は、前記第１の実施の形態のものと同一である。

本実施の形態において、頻出類似パターン検出手段２４は、類似構造生成手段２２から送出された部分構造の同値類の集合から、属性値の相違を無視しながら、頻出パターンの検出を行い、検出した頻出パターンを出力装置３に送出する。

図１０は、本発明の第２の実施形態に係る装置の動作を説明するための流れ図である。次に、図９及び図１０を参照して、本実施形態に係る装置の動作について詳細に説明する。本実施形態においては、図７のステップＡ３の代わりに、ステップＢ３が実行される。図１０のステップＡ１、Ａ２、Ａ４で示される処理は、前記第１の実施の形態における処理と同一であるため、説明は省略する。

前記第１の発明の実施の形態では、頻出パターン検出手段２３は、連結構造が同一でも属性値の異なる構造は同一と判定せずに、頻出パターンの検出を行っていた。

本実施の形態では、頻出類似パターン検出手段２４は、類似構造生成手段２２から与えられた同値類の集合を、連結構造が同一で属性値の異なる構造も同一な構造と判定しながら頻出パターンの検出を行い、検出された頻出パターンを出力装置３に送出する（図１０のステップＢ３）。

次に、本発明の第２の実施形態に係る装置の作用効果について説明する。

本発明の第２の実施の形態では、頻出類似パターン検出手段２４は、連結構造は

同一で属性値の異なる構造も同一な構造と判定しながら頻出パターンの検出を行うように構成されている。このため、意味は類似しているが属性値の異なる構造も同一な構造と判定して頻出パターンの検出を行うことができる。

次に、本発明の第３の実施形態について図面を参照して詳細に説明する。

図１１を参照すると、本発明の第３の実施の形態は、入力装置６を備え、データ処理装置５が類似構造生成調整手段２５及び類似構造判定調整手段２６を備えている以外は前記第２の実施の形態と同じである。

入力装置６は、使用者から、

・文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、

・属性値の種別ごとに値の差異を無視するか否かを指定するための入力と、

を受け付け、それぞれを、類似構造生成調整手段２５と類似構造判定調整手段２６に送出する。

入力装置６で受け付ける指定の入力の例としては、

・「使用者から文構造の差異の種別ごとに同一構造と判定するか否かと属性値の種別ごとに値の差異を無視するか否かについての指定項目」、

・「頻出パターン検出の際に同一パターンを持っていると判定しない文の例」、

・「頻出パターン検出の際に同一パターンを持っていると判定する文の例」

などが挙げられる。

類似構造生成調整手段２５は、入力装置６から与えられた指定から、連結構造の差異の種別ごとに同一構造と判定するか否かを決定し、その指定項目を、類似構造生成手段２２に送出する。

また、類似構造判定調整手段２６は、入力装置６から与えられた指定から、属性値の種別ごとに値の差異を無視するか否かを決定し、その指定項目を、頻出類似パターン検出手段２４に送出する。

類似構造生成手段２２は、類似構造生成調整手段２５からの指定に従って、言語解析手段２１より与えられた集合中の各構造の部分構造について、該部分構造の類似構造の生成を行い、その結果、生成された各類似構造を、それぞれの生成元の部分構造の同値類とする。

頻出類似パターン検出手段２４は、類似構造判定調整手段２６からの指定に従っ

て、属性値の差異の無視を行いながら、類似構造生成手段２２より与えられた同値類の集合から頻出パターンの検出を行う。

図１２は、本発明の第３の実施の形態に係る装置の動作を説明するための流れ図である。次に、図１１及び図１２のフローチャートを参照して本発明の第３の実施の形態に係る装置の動作について詳細に説明する。

最初に、言語解析手段２１がテキストＤＢ１１からテキスト集合を読み込む。

言語解析手段２１は、テキスト集合中の各テキストに対して解析を行い、解析結果として文構造を生成し、類似構造生成手段２２に送出する（図１２のステップＡ１）。図１２のステップＡ１における言語解析手段２１の動作は、前記第１の実施の形態における言語解析手段２１と同一である。

次に、入力装置６が、使用者から文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、属性値の種別ごとに値の差異を無視するか否かを指定するための入力とを受け付け、それぞれ類似構造生成調整手段２５と類似構造判定調整手段２６に送出する（図１２のステップＣ１）。

類似構造生成調整手段２５は、入力装置６からの指定を受け、文構造の差異の種別ごとに同一構造と判定するか否かの指定項目を生成し、類似構造生成手段２２に送出する。また、類似構造判定調整手段２６は、入力装置６からの指定を受け、属性値の種別ごとに値の差異を無視するか否かの指定項目を生成し頻出類似パターン検出手段２４に送出する（図１２のステップＣ２）。

類似構造生成手段２２は、類似構造生成調整手段２５からの指定に従って、言語解析手段２１より与えられた集合中の各文構造を構成する部分構造の類似構造の生成を行い、その結果、生成された各類似構造をそれぞれの生成元の部分構造の同値類とし、当該同値類の集合を頻出類似パターン検出手段２４に送出する（図１２のステップＣ３）。

頻出類似パターン検出手段２４は、類似構造判定調整手段２６からの指定に従って属性値の無視を行いながら、類似構造生成手段２２より与えられた同値類の集合から頻出パターンの検出を行う（図１２のステップＣ４）。

最後に、頻出類似パターン検出手段２４は、検出した頻出パターンを出力装置３に出力する（図１２のステップＡ４）。

図１３は、図１２のステップＣ３における、類似構造生成手段２２の動作の詳細なフローチャートである。

図１３を参照すると、類似構造生成手段２２は、

ステップＣ３－１の判定において、並列の変形が指定されている場合、並列の変形（図１３のステップＡ２－１）を行って部分構造の生成（図１３のステップＡ２－２）を行い、並列の変形が指定されていない場合、ステップＡ２－２の処理へ移行する。並列の変形、部分構造の生成は、図８のステップＡ２－１、Ａ２－２と同一である。

ステップＣ３－２の判定において、有向枝の無向枝化が指定されている場合、有向枝の無向枝化（図１３のステップＡ２－３）を行い、指定されていない場合、ステップＣ３－３の処理に移行する。有向枝の無向枝化は、図８のステップＡ２－３と同一である。

ステップＣ３－３の判定において、同義語の置換が指定されている場合、同義語の置換（図１３のステップＡ２－４）を行い、同義語の置換が指定されていない場合、ステップＣ３－４の処理に進む。同義語の置換は、図８のステップＡ２－４と同一である。

ステップＣ３－３の判定において、順序木の無順序木化が指定されている場合、順序木の無順序木化（図１３のステップＡ２－５）を行い、指定されていない場合、ステップＡ２－６の処理に移行する。

ステップＡ２－６では、同値類を生成する。順序木の無順序木化、同値類を生成は、図８のステップＡ２－５、Ａ２－６と同一である。

このように、本実施の形態では、並列の変形（図１３のステップＡ２－１）、有向枝の無向枝化（図１３のステップＡ２－３）、同義語の置換（図１３のステップＡ２－４）、及び、順序木の無順序木化（図１３のステップＡ２－５）が、類似構造生成調整手段２５から与えられた指定により、実行の有無が制御される点で、図８に示した前記第１の実施の形態の類似構造生成手段２２と相違している。

使用者は、出力されたパターンを参照して、ステップＣ１に戻りどこまで類似した構造を同一と判定するかを指定するための入力を再度行ったうえで本発明に頻出パターン検出を再度行わせることができる。

次に、本発明の第３の実施の形態に係る装置の作用効果について説明する。

本実施の形態では、類似構造生成調整手段と類似構造判定調整手段が使用者からの指定に基づきどこまで類似した構造を同一な構造と判定するかの調整を行うように構成されている。このため、使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整できる。

次に、本発明の第４の実施の形態について図面を参照して詳細に説明する。

図１４を参照すると、本発明の第４の実施の形態に係る装置は、前記した第１、第２、第３の実施の形態をプログラムにより構成したものである。図１４はこの場合に、そのプログラムにより動作されるコンピュータの構成を示す図である。

テキストマイニング用プログラム７は、データ処理装置８に読み込まれ、データ処理装置８の動作を制御する。データ処理装置８はテキストマイニング用プログラム７の制御により以下の処理、すなわち第１、第２及び第３の実施の形態におけるデータ処理装置２、４及び５による処理と同一の処理を実行する。

次に、本発明を具体的な実施例を即して詳細に説明する。

まず、本発明の第１の実施例について図面を参照して説明する。本発明の第１の実施例は前記第１の実施の形態の一具体例である。

本実施例における装置は、図６のデータ処理装置２をパーソナル・コンピュータで、記憶装置１を磁気ディスク記憶装置で、出力装置３としてディスプレイを備えて構成されている。

パーソナル・コンピュータ２は、言語解析手段２１、類似構造生成手段２２、頻出パターン検出手段２３として機能する中央演算装置（ＣＰＵ）を有している。磁気ディスク記憶装置には、テキストＤＢ１１としてテキスト集合が記憶されている。

図１５は、テキスト集合の内容を示す図である。

言語解析手段２１は、テキストＤＢ１１中の図１５に示されるテキスト集合の各テキストに対して言語解析を行い、その結果、各テキストの文構造を得る（図７のステップＡ１）。

図１６Ａ～図１６Ｃに、それぞれ言語解析手段２１で得られる文１～文３の文構造を示す。

次に、類似構造生成手段２２は、図１６Ａ～図１６Ｃに示される各文構造を構成

する部分構造の全ての類似構造を生成し、その結果、生成された類似構造を生成元の部分構造の同値類とする（図７のステップＡ２）。

本実施例では、図１６Ｂに示される文２（「速く安い車種Ａ」）の文構造から、部分構造の同値類を生成する様子を例にとって説明する。この例は、図１８～２１に示されている。

類似構造生成手段２２は、まず、図１８に示すように、並列構造の変形を行い（図８のステップＡ２－１）、次に部分構造２ａ－０において、並列関係にある「速い」と「安い」の接続関係を変形し、類似構造２ａ－１を生成する。

類似構造生成手段２２は、次に、図１９に示すように、部分構造の生成を行い（図８のステップＡ２－２）、部分構造２ａ－０から、２単語の関係を表す部分構造２ｃ－０及び２ｇ－０と、１単語の部分構造２ｄ－０、２ｅ－０及び２ｆ－０を生成する。

類似構造生成手段２２は、また、類似構造２ａ－１から、部分構造２ａ－０に含まれない２単語の関係を表す部分構造２ｂ－０を生成する。

なお、部分構造２ａ－０と類似構造２ａ－１の両方から生成される構造は１つにまとめて扱う。

また、ここで部分構造を生成するのに用いた部分構造２ａ－０、及び類似構造２ａ－１も、今後の類似構造生成において、部分構造及び類似構造として扱う。

次に、類似構造生成手段２２は有向枝の無向枝化を行う（図８のステップＡ２－３）。この例においては、ステップＡ２－２において生成した部分構造の全ての有向枝が無向枝化され、新たな類似構造が生成される。図２０Ａに示すように、例えば部分構造２ａ－０の有向枝を無向枝化して、類似構造２ａ－２が生成される。なお、１単語からなり有向枝を持たない部分構造２ｄ－０、２ｅ－０及び２ｆ－０は、ステップＡ２－３では変形が行われないため、図２０Ａ～図２０Ｅでは省略されている。

次に、同義語の置換が行われる（図８のステップＡ２－４）。本実施例における「同義語の置換」では、ユーザによりあらかじめ与えられた同義語辞書に定義された被置換語を代表語に置き換えるものとする。

また、本実施例に用いる同義語辞書は、図１７に示されるように、被置換語「高

速」を代表語「速い」に置き換える１つの辞書項目のみが登録された同義語辞書が指定されたものとしている。

この時点で生成された部分構造及び類似構造には、被置換語「高速」が含まれないため、ステップＡ２－４では、変形が発生しない。そのため、ここではステップＡ２－４による変形の図を省略している。

次に、順序木の無順序木化が行われる（図８のステップＡ２－５）。ここでは、文構造の木構造において、兄弟関係にある単語を５０音順にソートすることによって、順序木の無順序木化を行う。

なお、順序木の無順序木化を行うための他の方法として、

・兄弟関係にある単語を５０音順以外の一定の法則に従いソートする方法や、

・ソートを行わずに頻出類似パターン検出時に兄弟関係にある単語の順序だけが異なる木を同一と判定する方法を用いてもよい。

生成された部分構造及び類似構造では、類似構造２ａ－１及び２ａ－３（図２０Ｃ）を除いた部分構造及び類似構造では兄弟関係になっている単語が存在しない。類似構造２ａ－１及び２ａ－３では、既に兄弟関係にある単語が５０音順に並んでいる。このため、実質的に変形が発生しない。そのため、ここではステップＡ２－５による変形の図を省略している。

最後に、類似構造を生成元の部分構造の同値類とすることで、同値類の生成が行われる（図８のステップＡ２－６）。

図２０Ａ～図２０Ｅに示された部分構造及び類似構造の集合において、各類似構造を生成元の部分構造の同値類することで生成される同値類を図２１に示す。部分構造２ａ－０と、部分構造２ａ－０の有向枝を無向枝化することで生成された類似構造２ａ－２と、部分構造２ａ－０を並列変形した類似構造２ａ－１と、類似構造２ａ－１の有向枝を無向枝化することで生成された類似構造２ａ－３とは同値類２ａを構成している。

部分構造２ｂ－０と、部分構造２ｂ－０の有向枝を無向枝化することで生成された類似構造２ｂ－１は、同値類２ｂを構成している。部分構造２ｃ－０と、部分構造２ｃ－０の有向枝を無向枝化することによって生成された類似構造２ｃ－１は、同値類２ｃを構成している。部分構造２ｇ－０と、部分構造２ｇ－０の有向枝を無

向枝化することによって生成された類似構造２ｇ－１は、同値類２ｇを構成している。部分構造２ｄ－０、２ｅ－０、２ｆ－０は、類似構造と部分構造は同一である。

図１８～図２１に示したように、本実施例において、文２の文構造（図１６Ｂ参照）から、類似構造生成手段２２が同値類を生成する例においては、同義語の置換（図８のステップＡ２－４）及び順序木の無順序木化（図８のステップＡ２－５）で変形は行われない。

図２２に示すように、文３の文構造（図１６Ｃ参照）を構成する一の部分構造に対して、類似構造生成手段２２による変形処理が行われる。以下、同義語の置換（図８のステップＡ２－４）及び順序木の無順序木化（図８のステップＡ２－５）で発生する変形の例を説明する。

まず文３の文構造を表す部分構造３ａ－０に対して並列の変形（図８のステップＡ２－１）が行われる。ここでは、部分構造３ａ－０が並列の構造を含まず変形が行われないため、図２２には、並列の変形による結果の構造は含まれない。

次に、部分構造３ａ－０から部分構造の生成（図８のステップＡ２－２）が行われる。ここでは、部分構造３ａ－０に行われる構造変形にのみ注目して説明するため、部分構造３ａ－０から、他の部分構造を生成する処理である部分構造の生成は省略する。

次に、部分構造３ａ－０に対して、有向枝の無向枝化（図８のステップＡ２－３）が行われる。部分構造３ａ－０の「安い」から、「車種Ａ」への有向枝と、「高速」から「車種Ａ」への有向枝が無向枝化される。その結果、類似構造３ａ－１が生成される（図２２：ステップＡ２－３）。

次に、類似構造３ａ－１に対して、同義語の置換（図８のステップＡ２－４）が行われる。ここでは、図１７に示される同義語辞書を用いているため、被置換語「高速」が代表語「速い」に置き換えられる。類似構造３ａ－１に含まれる被置換語「高速」も代表語「速い」に置き換えられ、類似構造に変形される（図２２：ステップＡ２－４）。

次に、類似構造３ａ－１に対して、順序木の無順序木化（図８のステップＡ２－５）が行われる。ここでは、兄弟関係にある単語を、５０音順にソートすることで順序木の無順序木化が行われる。このため、類似構造３ａ－１において兄弟関係に

ある「安い」と「速い」の順序を入れ替え、５０音順にソートにされ、類似構造に変換される（図２２：ステップＡ２－５）。

このようにして生成された類似構造に対して同値類の生成（図８のステップＡ２－６）が行われる。尚、本実施例では、部分構造３ａ－０から生成される一つの類似構造３ａ－１に行われる変形のみに注目して説明しているためその説明を省略する。

このようにして、類似構造生成手段２２が、部分構造と類似構造及び同値類の生成を行うことで、本実施例では、図１６Ａの文１の文構造から、図２３に示すような同値類が生成される。図１６Ｂの文２の文構造から、図２４に示すような同値類が生成される。また図１６Ｃの文３の文構造から、図２５に示すような同値類が生成される。

ただし、本来は、図２２における変形の途中経過（図２２：ステップＡ２－３からステップＡ２－４における類似構造３ａ－１）のように、形の違う類似構造も生成されている。尚、説明を分かりやすくするため、頻出パターンの検出に用いられない構造は、図２３～図２５の同値類からは省略している。

次に、頻出パターン検出手段２３は、図２３～図２５に示される同値類の集合から頻出パターン（頻出する同値類）の検出を行う（図７のステップＡ３）。

この際、頻出パターン検出手段２３は、要素の少なくとも一つが同一である同値類は、同一と判定して、頻出パターンの検出を行う。

例えば、本実施例においては、図２３の同値類１ｃの要素である類似構造１ｃ－１と、図２４の同値類２ｂの要素である類似構造２ｂ－１は、どちらも「車種Ａ」と「速い」が無向枝で連結された構造で、属性値の差分もないため、同一の構造である。

従って、頻出パターン検出手段２３は、図２３の同値類１ｃと図２４の同値類２ｂを同一と判定する。

図２３～図２５を参照すると、

「類似構造１ｃ－１、類似構造２ｂ－１と、類似構造３ｃ－１」、

「部分構造１ｄ－０、部分構造２ｄ－０と、類似構造３ｅ－１」、

「部分構造１ｅ－０、部分構造２ｆ－０と、部分構造３ｆ－０」、

「部分構造１ｆ－０と部分構造２ｅ－０」

がそれぞれ同一の構造となっている。

「要素の少なくとも一つが同一である同値類は同一と判定する」という同値類の性質により、図２３～図２５に示される同値類のうち、

「同値類１ｃ、２ｂ、及び、３ｃ」、

「同値類１ｄ、２ｄ、及び、３ｅ」、

「同値類１ｅ、２ｆ、及び、３ｆ」、

「同値類１ｆ、及び、２ｅ」

がそれぞれ同一の同値類と判定される。

本実施例では、３回以上出現する同値類を頻出パターンとする。なお、どのような出現回数の同値類を頻出パターンとして検出するかは、使用者がテキストマイニングを実行する前に決定することができる。

この場合、

「同値類１ｃ、２ｂ、及び、３ｃ」、

「同値類１ｄ、２ｄ、及び、３ｅ」、

「同値類１ｅ、２ｆ、及び、３ｆ」

が頻出パターンとして検出される。

最後に、そのようにして抽出された頻出パターンを表す構造を出力装置３に表示する（図７のステップＡ４）。

図２６は、本実施例において、出力装置３が出力する頻出パターンの表現の一例を示す図である。本実施例では、頻出パターンを表す同値類の要素である類似構造を、頻出パターンの表現として用いている。

類似構造を生成し、同値類を生成して頻出パターンの検出を行うことによって、「部分構造１ｃ－０（図２３）、部分構造２ｂ－０（図２４）、及び、部分構造３ｃ－０（図２５）」のように類似した意味を持つが、連結構造の異なる部分構造を同一と判定し、頻出パターンとして検出することができる。

次に、本発明の第２の実施例について図面を参照して説明する。本実施例は、前記第２の実施の形態に対応するものである。

本実施例に係る装置は、データ処理装置４をパーソナル・コンピュータで、記憶

装置１を磁気ディスク記憶装置で、出力装置３としてディスプレイを備えて構成されている。

パーソナル・コンピュータ４は、言語解析手段２１、類似構造生成手段２２、頻出類似パターン検出手段２４として機能する中央演算装置（ＣＰＵ）を有し、磁気ディスク記憶装置には、テキストＤＢ１１としてテキスト集合が記憶されている。テキスト集合としては、前記第１の実施例と同様、図１５に示した文１～文３を使用する。

言語解析手段２１は、テキストＤＢ１１中の図１５に示されるテキスト集合の各テキストに対して、言語解析を行い、各テキストの文構造を得る（図１０のステップＡ１）。ここで得られる文構造は、前記第１の実施例と同様、図１６Ａ～図１６Ｃのようになる。

次に、類似構造生成手段２２は、図１６Ａ～図１６Ｃに示される各文構造を構成する部分構造の全ての類似構造を生成し、その結果、生成された類似構造を生成元の部分構造の同値類とする（図１０のステップＡ２）。ここで得られる同値類は、前記第１の実施例と同様、図２３～図２５のようになる。

次に、頻出類似パターン検出手段２４は、図２３～図２５に示される同値類の集合から、属性値の差異を無視しながら頻出パターン（頻出する同値類）の検出を行う（図１０のステップＢ３）。

頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定して、頻出パターンの検出を行う。ただし、本実施例の頻出類似パターン検出手段２４は、表層格や付属語情報などの属性値の差異を無視して、類似構造の同一性の判定を行っており、この点で、前記第１の実施例の頻出パターン検出手段２３と相違している。

例えば、図２３の類似構造１ａ－１と図２４の類似構造２ａ－３は、どちらも、「車種Ａ」と「速い」及び「安い」が無向枝で連結された構造である。しかし、表層格が異なるため、前記第１の実施例の頻出パターン検出手段２３では、同一と判定されない。一方、本実施例の頻出類似パターン検出手段２４では、同一と判定される。

本実施例においては、図２３～図２５を参照すると、

「類似構造１ａ－１、類似構造２ａ－３、及び、類似構造３ａ－１」、

「類似構造１ｂ－１、類似構造２ｃ－１と、類似構造３ｂ－１」、

「類似構造１ｃ－１、類似構造２ｂ－１、及び、類似構造３ｃ－１」、

「部分構造１ｄ－０、部分構造２ｄ－０、及び、類似構造３ｅ－１」、

「部分構造１ｅ－０、部分構造２ｆ－０、及び、部分構造３ｆ－０」、

「部分構造１ｆ－０、部分構造２ｅ－０、及び、部分構造３ｄ－０」

がそれぞれ頻出類似パターン検出手段２４に同一の構造と判定される。

頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定するため、

「同値類１ａ、２ａ、及び、３ａ」、

「同値類１ｂ、２ｃ、及び、３ｂ」、

「同値類１ｃ、２ｂ、及び、３ｃ」、

「同値類１ｄ、２ｄ、及び、３ｅ」、

「同値類１ｅ、２ｆ、及び、３ｆ」、

「同値類１ｆ、２ｅ、及び、３ｄ」

をそれぞれ同一の同値類と判定する。

本実施例では、前記第１の実施例と同様に、３回以上出現する同値類を頻出パターンとする。この場合、

「同値類１ａ、２ａ、及び、３ａ」、

「同値類１ｂ、２ｃ、及び、３ｂ」、

「同値類１ｃ、２ｂ、及び、３ｃ」、

「同値類１ｄ、２ｄ、及び、３ｅ」、

「同値類１ｅ、２ｆ、及び、３ｆ」、

「同値類１ｆ、２ｅ、及び、３ｄ」

が頻出パターンとして検出される。

最後に、そのようにして抽出された頻出パターンを表す構造を、出力装置３に表示する（図１０のステップＡ４）。

本実施例において、出力装置３が出力する頻出パターンの表現は図２７のようになる。本実施例では、前記第１の実施例と同様に、頻出パターンを表す同値類の要

素である類似構造を、頻出パターンの表現として用いている。

このようにして、属性値の差異を無視して頻出パターンの検出を行うことによって、

「部分構造１ｂ－０（図２３）、部分構造２ｃ－０（図２４）と部分構造３ｂ－０（図２５）」、

「部分構造１ｆ－０（図２３）、部分構造２ｅ－０（図２４）と部分構造３ｆ－０（図２５）」

のように、類似した意味を持つが属性値の異なる部分構造を同一と判定し、頻出パターンとして、検出を行うことができる。

次に、本発明の第３の実施例について図面を参照して説明する。本実施例は、本発明の第３の実施の形態に対応するものである。

本実施例に係る装置は、データ処理装置５をパーソナル・コンピュータで、記憶装置１を磁気ディスク記憶装置で、出力装置３としてディスプレイを、入力装置６としてキーボードを備えて構成されている。

パーソナル・コンピュータ５は、言語解析手段２１、類似構造生成手段２２、頻出類似パターン検出手段２４、類似構造生成調整手段２５、類似構造判定調整手段２６として機能する中央演算装置（ＣＰＵ）を有している。磁気ディスク記憶装置には、テキストＤＢ１１としてテキスト集合が記憶されている。テキスト集合としては、前記第１、第２の実施例と同様、図１５に示した文が用いられる。

言語解析装置２１は、テキストＤＢ１１中の図１５に示されるテキスト集合の各テキストに対して、言語解析を行い、各テキストの文構造を得る（図１２のステップＡ１）。ここで得られる文構造は、前記第１、第２の実施例と同じく、図１６Ａ～図１６Ｃのようになる。

次に、使用者は、入力装置６を用いて、

・文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、

・属性値の種別ごとに値の差異を無視するか否かを指定するための入力

を行う（図１２のステップＣ１）。

本実施例において、例えば

「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は

同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異は同一と判定する」という入力を行ったとする。

入力装置６は、使用者から受け付けた入力を、類似構造生成調整手段２５と類似構造判定調整手段２６に送出する。

次に、類似構造生成調整手段２５は、入力装置６から使用者の指定を受け取り、類似構造生成手段２２の動作を制御する（図１２のステップＣ２）。

本実施例においては、類似構造生成調整手段２５は、入力装置６から、

「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異は同一と判定する」

という指定を受け取ると、

類似構造生成手段２２が行う、文構造の部分構造から類似構造を生成する際の変形処理、すなわち並列構造の変形（図１３のステップＡ２－１）、有向枝の無向枝化（図１３のステップＡ２－３）及び順序木の無順序木化（図１３のステップＡ２－５）は行われる。しかし、類似構造生成調整手段２５は、同義語の置換（図１３のステップＡ２－４）がスキップされるように類似構造生成手段２２の動作を制御する。

一方、類似構造判定調整手段２６は、入力装置６から使用者の入力を受け取り、頻出類似パターン検出手段２４の動作を制御する（図１２のステップＣ２）。

本実施例においては、類似構造生成調整手段２６は、入力装置６から、「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。

属性値の差異については、類似構造生成調整手段２６は、付属語情報の差異と表層格の差異については同一と判定する」という指定を受け取り、頻出類似パターン検出手段２４が類似構造の同一性判定の処理を、表層格の差異及び付属語情報の差異を無視して行うように制御する。

次に、類似構造生成手段２２は、図１６Ａ～図１６Ｃに示される各文構造の部分構造についてステップＣ２で生成した指定項目に従い、同義語の置換（図１３のス

テップＡ２－４）を飛ばして類似構造を生成し、その結果、生成された類似構造を生成元の部分構造の同値類とする（図１２のステップＣ３）。

以下、図１６Ｃに示される文３の文構造の一部分構造に対して、類似構造生成手段２２が行う変形を例にとって説明する。図２８に、その一例を示す。

まず、文３の文構造を表す部分構造３ａ－０に対して、並列の変形（図１３のステップＡ２－１）が行われる。ただし、図２８に示す例では、部分構造３ａ－０が並列の構造を含まず変形が行われないため、図２８には、並列の変形による結果の構造は含まれない。

次に、部分構造３ａ－０から部分構造の生成（図１３のステップＡ２－２）が行われる。尚、部分構造３ａ－０に行われる構造変形にのみ注目して説明するため、部分構造３ａ－０から他の部分構造を生成する処理である部分構造の生成は省略する。

次に、部分構造３ａ－０に対して有向枝の無向枝化（図１３のステップＡ２－３）が行われる。部分構造３ａ－０の「安い」から「車種Ａ」への有向枝と、「高速」から「車種Ａ」への有向枝が無向枝化される。その結果、類似構造３ａ－２が生成される（図２８のステップＡ２－３）。

同義語の置換（図１３のステップＡ２－４）は、類似構造生成調整手段２５より与えられた指定により、ステップＣ３－３の判定でスキップされるため、実行されない。

次に、類似構造３ａ－２に対して、順序木の無順序木化（図１３のステップＡ２－５）が行われる。ここでは、兄弟関係にある単語を５０音順にソートすることで、順序木の無順序木化が行われる。類似構造３ａ－２（図２８のステップＡ２－３処理後における類似構造）において、兄弟関係にある単語「安い」と「高速」の順序を入れ替えるように、前記単語が５０音順にソートされる。その結果、類似構造３ａ－２は図２８のステップＡ２－５処理後における類似構造に変換される。

このようにして生成された類似構造に対して、同値類の生成（図１３のステップＡ２－６）が行われる。尚、部分構造３ａ－０から生成される一つの類似構造３ａ－２に行われる変形のみに注目して説明しているため、省略する。

本実施例における変形では、同義語の置換（図１３のステップＡ２－４）が飛ば

されるため、図２８のステップＡ２－５処理後における類似構造３ａ－２には、被置換語「高速」が残っている。一方、図２２に示した前記第１、第２の実施例における変形の例、すなわちステップＡ２－５処理後における類似構造３ａ－１では、被置換語「高速」が代表語「速い」に置換されている。

本実施例では、このようにして、類似構造生成手段２２が部分構造と類似構造及び同値類の生成を行うことで、図１６Ａに示される文１の文構造から、図２３に示される同値類が生成され、図１６Ｂに示される文２の文構造から、図２４に示される同値類が生成され、図１６Ｃに示される文３の文構造から図２９に示される同値類が生成される。

次に、頻出類似パターン検出手段２４は、図２３、図２４、及び図２９に示される同値類の集合から、ステップＣ２で、類似構造判定調整手段２６が指定した属性値の差異を無視しながら頻出パターンの検出を行う（図１２ のステップＣ４）。

頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定して、頻出パターンの検出を行う。

本実施例においては、頻出類似パターン検出手段２４は、類似構造判定調整手段２６からの指定により、どの属性値の差異を無視して類似構造の同一性を判定するかを決定する。

本実施例では、

「表層格の差異を無視する」、

「付属語情報の差異を無視する」

と動作を制御するように類似構造判定調整手段２６が指定を行ったため、頻出類似パターン検出手段２４は、前記第２の実施例と同様に、類似構造の同一性の判定を行う。

本実施例においては、図２３、図２４、及び図２９を参照すると、

「類似構造１ａ－１、及び、類似構造２ａ－３」、

「部分構造２ｃ－０、及び、部分構造３ｂ－０」、

「類似構造１ｂ－１、類似構造２ｃ－１、及び、類似構造３ｂ－１」、

「部分構造１ｃ－０、及び、類似構造２ｂ－０」、

「類似構造１ｃ－１、及び、類似構造２ｂ－１」、

「部分構造１ｄ－０、及び、部分構造２ｄ－０」、

「部分構造１ｅ－０、部分構造２ｆ－０、及び、部分構造３ｆ－０」、

「部分構造１ｆ－０、部分構造２ｅ－０、及び、部分構造３ｄ－０」

がそれぞれ頻出類似パターン検出手段２４に同一の構造と判定される。

頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定するため、

「同値類１ａ、及び、２ａ」、

「同値類１ｂ、２ｃ、及び、３ｂ」、

「同値類１ｃ、及び、２ｂ」、

「同値類１ｄ、及び、２ｄ」、

「同値類１ｅ、２ｆ、及び、３ｆ」、

「同値類１ｆ、２ｅ、及び、３ｄ」

をそれぞれ同一の同値類と判定する。

本実施例では、前記第１、第２の実施例と同様に、３回以上出現する同値類を頻出パターンとする。

この場合、

「同値類１ｂ、２ｃ、及び、３ｂ」、

「同値類１ｅ、２ｆ、及び、３ｆ」、

「同値類１ｆ、２ｅ、及び、３ｄ」

が頻出パターンとして検出される。

最後に、このようにして抽出された頻出パターンを表す構造を、出力装置３に表示する（図１２のステップＡ４）。

本実施例において、出力装置３が出力する頻出パターンの表現は、図３０のようになる。図３０に示すように、本実施例では、前記第１、第２の実施例と同様に、頻出パターンを表す同値類の要素である類似構造を頻出パターンの表現として用いている。

使用者は、この頻出パターン検出に不満を感じた場合、図１２のステップＣ１に戻り、どこまで類似した構造を同一と判定するかの指定の入力を変更することで、再度頻出パターンの検出を行うことができる。

このようにして、

「同義語の置換による差異については同一と判定しない」

という使用者の指定に基づき、図２３、図２４、図２９において、

「部分構造１ａ－０、部分構造２ａ－０、及び、部分構造３ａ－０」、

「部分構造１ｃ－０、部分構造２ｂ－０、及び、部分構造３ｃ－０」、

「部分構造１ｄ－０、部分構造２ｄ－０、及び、部分構造３ｅ－０」

といった類似した意味を持つが使用者の入力に反する構造を同一と判定せずに、頻出パターン検出行うことで、使用者がどこまで類似した構造を同一と判定するかの調整を行うことができる。

本発明によれば、連結構造は異なるが類似した意味を持つ構造を同一の構造と判定して頻出パターンを検出することができる。本発明によれば、属性値を持たない構造の集合に対して類似構造を同一と判定して頻出パターンの検出を行うことができる。

その理由は、本発明においては、生成した類似構造を元の構造の同値類として扱い、頻出パターン検出を行う構成としたためである。本発明によれば、属性値を持つ構造の集合に対しても類似構造を同一と判定して頻出パターンの検出を行うことができる。

また、本発明によれば、類似した意味を持つが異なる属性値を持つ構造を同一の構造と判定して頻出パターンを検出することができる。

その理由は、本発明においては、頻出類似パターン検出手段が属性値の差異を無視して頻出パターン検出を行うためである。

さらに本発明によれば、テキストマイニング装置の使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整することができる。

その理由は、本発明においては、類似構造生成調整手段と類似構造判定調整手段が使用者からの入力に基づき、どこまで類似した構造を同一な構造と判定するかの調整を行う構成としたためである。

産業上の利用可能性

本発明によれば、コンピュータ上に蓄積される、顧客からの苦情メールやアンケ

ート結果の特徴分析を行う目的に良く用いられるテキストマイニング装置や、テキストマイニング装置をコンピュータに実現するためのプログラムといった用途に適用できる。

請 求 の 範 囲

１． 入力した文書から文構造を作成する手段と、

前記文構造の部分構造に対して、予め定められた所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する手段と、

前記意味の類似したパターンを同一パターンと判定してパターン検出を行う手段と、

を備えていることを特徴とするテキストマイニング装置。

２． テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部の前記文書を入力して解析し文構造を取得する解析部と、

を備え、

前記解析部は、前記文書を解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する、

ことを特徴とする請求項１に記載のテキストマイニング装置。

３． 前記類似構造を生成する手段が、

前記文構造について並列変形を行う手段と、

前記文構造の部分構成を生成する手段と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、

を備え、

前記類似構造を前記文構造の部分構造の同値類とする、

ことを特徴とする、請求項１に記載のテキストマイニング装置。

４． テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、

前記解析部により解析して得られる文構造の部分構造に対して予め定められた所定の変形操作を行い、意味的に類似したパターンの類似構造を生成する類似構造生成部と、

前記類似構造生成部によって生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行うパターン検出部と、

を備えていることを特徴とするテキストマイニング装置。

５． 前記パターン検出部は、前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出することを特徴とする請求項４に記載のテキストマイニング装置。

６． 前記類似構造生成部が、

前記文構造について並列変形を行う手段と、

前記文構造の部分構成を生成する手段と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、

を備え、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、

ことを特徴とする請求項４に記載のテキストマイニング装置。

７． 使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する手段を備えていることを特徴とする請求項４に記載のテキストマイニング装置。

８． テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、

使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目を生成する類似構造生成調整部と、

使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成する類似構造判定調整部と、

前記類似構造生成調整部によって生成された第１の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する類似構造生成部と、

前記類似構造生成部によって生成された類似構造を生成元の部分構造の同値類として扱い、前記類似構造判定調整部の第２の指定項目に従い、属性値の差異を無視しながら、頻出パターンの検出を行う類似パターン検出部と、

を備えていることを特徴とするテキストマイニング装置。

９．　前記解析部は、前記文書を解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした前記文構造を生成し、

前記属性値は、前記文構造に付加された表層格及び／又は付属語情報を含む、

ことを特徴とする請求項８に記載のテキストマイニング装置。

１０．　前記類似パターン検出部は、頻出の類似パターンを検出することを特徴とする請求項８に記載のテキストマイニング装置。

１１．　前記類似構造生成部が、

前記第１の指定項目に、並列変形の指定がある場合、前記文構造について並列変形を行う手段と、

前記文構造の部分構造を生成する手段と、

前記第１の指定項目に、有向枝の無向枝化の指定がある場合に、前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、

前記第１の指定項目に、同義語の置換の指定がある場合、同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、

前記第１の指定項目に、順序木の無順序木化の指定がある場合、前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、

を備え、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、

ことを特徴とする請求項８に記載のテキストマイニング装置。

１２．　入力した文書から文構造を作成する工程と、

前記文構造の部分構造に対する所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する工程と、

前記意味の類似したパターンを同一パターンと判定してパターン検出を行う工程と、

を含むことを特徴とするテキストマイニング方法。

１３．　テキストマイニングの対象となる文書の集まりを記憶する記憶部から前記文書を入力して解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する工程を含むことを特徴とする請求項１２に記載のテキストマイニング方法。

１４．　前記類似構造を生成する工程が、

前記文構造について並列変形を行う工程と、

前記文構造の部分構造を生成する工程と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、

を含み、

前記類似構造を前記部分構造の同値類とする、

ことを特徴とする請求項１２に記載のテキストマイニング方法。

１５．　テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、

前記文構造の部分構造に対して予め定められた所定の変形操作を行い、意味的に類似したパターンを有する類似構造を生成する工程と、

前記生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行う工程と、

を含むことを特徴とするテキストマイニング方法。

１６．　前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出する工程を含むことを特徴とする請求項１５に記載のテキストマイニング方法。

１７．　前記類似構造を生成する工程が、

前記文構造について並列変形を行う工程と、

前記文構造の部分構造を生成する工程と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、

を含み、

前記文構造の類似構造を生成し、前記類似構造を同値類とすること、

を特徴とする請求項１５に記載のテキストマイニング方法。

１８．　使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する工程を備えていることを特徴とする請求項１７に記載のテキストマイニング方法。

１９．　テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、

使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目を生成する工程と、

使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成する工程と、

前記生成された第１の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変形操作を行い、前記部分構造と意味的に類似した類似構造を生成する工程と、

前記生成された類似構造を生成元の部分構造の同値類として扱い、前記第２の指定項目に従い、属性値の差異を無視してパターンの検出を行う工程と、

を含むことを特徴とするテキストマイニング方法。

２０．　前記文構造を取得する工程が、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした前記文構造を生成し、

前記属性値は、前記文構造に付加された表層格及び／又は付属語情報を含む、

ことを特徴とする請求項１９に記載のテキストマイニング方法。

２１．　前記頻出の類似パターンを検出することを特徴とする請求項１９に記載のテキストマイニング方法。

２２．　前記類似構造を生成する工程が、

前記第１の指定項目に、並列変形の指定がある場合、前記文構造について並列変形を行う工程と、

前記文構造の部分構造を生成する工程と、

前記第１の指定項目に、有向枝の無向枝化の指定がある場合に、前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

前記第１の指定項目に、同義語の置換の指定がある場合、同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、

前記第１の指定項目に、順序木の無順序木化の指定がある場合、前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、

を含み、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項１９に記載のテキストマイニング方法。

２３．　テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、

前記文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する処理と、

前記生成された類似構造を、生成元の部分構造の同値類として扱い、所定のパターン検出を行う処理と、

を実行させるプログラム。

２４．　テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、

前記文構造の部分構造に対して予め定められた所定の変換操作を行い、前記部分構造と意味的に類似したパターンの類似構造を生成する処理と、

前記生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異を無視しながらパターン検出を行う処理と、

を実行させるプログラム。

２５．　テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、

使用者の入力から、前記文構造の差異の種別ごとに、同一構造と判定するか否かを指定する第１の指定項目と、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成する処理と、

前記文構造の差異の種別ごとに同一構造と判定するか否かを指定する前記第１

の指定項目に従って、前記文構造の部分構造に対して所定の変換操作を行い、意味的に類似したパターンの類似構造を生成する処理と、

生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異の種別ごとに同一構造と判定するか否かを指定する前記第２の指定項目に従って、属性値の差異を無視しながら頻出パターンの検出を行う処理と

を実行させるプログラム。

補正書の請求の範囲

［２００５年７月１５日（１５．０７．２００５）国際事務局受理：出願当初の請求の範囲１、４、１２、１５、２３及び２４は補正された；他の請求の範囲は変更なし。（５頁）］

１．（補正後）　入力した文書から文構造を作成する手段と、

前記文構造の部分構造に対して、少なくともグラフ構造の枝の繋ぎ変えを含む、予め定められた所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を生成する手段と、

前記意味の類似したパターンを同一パターンと判定してパターン検出を行う手段と、

を備えていることを特徴とするテキストマイニング装置。

２．　テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部の前記文書を入力して解析し文構造を取得する解析部と、

を備え、

前記解析部は、前記文書を解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する、

ことを特徴とする請求項１に記載のテキストマイニング装置。

３．　前記類似構造を生成する手段が、

前記文構造について並列変形を行う手段と、

前記文構造の部分構成を生成する手段と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、

を備え、

前記類似構造を前記文構造の部分構造の同値類とする、

ことを特徴とする、請求項１に記載のテキストマイニング装置。

４．（補正後）テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、

前記解析部により解析して得られる文構造の部分構造に対して、少なくともグラフ構造の枝の繋ぎ変えを含む、予め定められた所定の変換操作を行い、意味的に類似したパターンの類似構造を生成する類似構造生成部と、

前記類似構造生成部によって生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行うパターン検出部と、

を備えていることを特徴とするテキストマイニング装置。

５． 前記パターン検出部は、前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出することを特徴とする請求項４に記載のテキストマイニング装置。

６． 前記類似構造生成部が、

前記文構造について並列変形を行う手段と、

前記文構造の部分構成を生成する手段と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、

を備え、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、

ことを特徴とする請求項４に記載のテキストマイニング装置。

７． 使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する手段を備えていることを特徴とする請求項４に記載のテキストマイニング装置。

８． テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、

前記第１の指定項目に、順序木の無順序木化の指定がある場合、前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、

を備え、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、

ことを特徴とする請求項８に記載のテキストマイニング装置。

１２．（補正後）　入力した文書から文構造を作成する工程と、

前記文構造の部分構造に対して、少なくともグラフ構造の枝の繋ぎ変えを含む、所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する工程と、

前記意味の類似したパターンを同一パターンと判定してパターン検出を行う工程と、

を含むことを特徴とするテキストマイニング方法。

１３．　テキストマイニングの対象となる文書の集まりを記憶する記憶部から前記文書を入力して解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する工程を含むことを特徴とする請求項１２に記載のテキストマイニング方法。

１４．　前記類似構造を生成する工程が、

前記文構造について並列変形を行う工程と、

前記文構造の部分構造を生成する工程と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、

を含み、

前記類似構造を前記部分構造の同値類とする、

ことを特徴とする請求項１２に記載のテキストマイニング方法。

１５．（補正後）　テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、

前記文構造の部分構造に対して、少なくともグラフ構造の枝の繋ぎ変えを含む、予め定められた所定の変換操作を行い、意味的に類似したパターンを有する類似構造を生成する工程と、

前記生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行う工程と、

を含むことを特徴とするテキストマイニング方法。

１６．　前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出する工程を含むことを特徴とする請求項１５に記載のテキストマイニング方法。

１７．　前記類似構造を生成する工程が、

前記文構造について並列変形を行う工程と、

前記文構造の部分構造を生成する工程と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、

を含み、

前記文構造の類似構造を生成し、前記類似構造を同値類とすること、

を特徴とする請求項１５に記載のテキストマイニング方法。

１８．　使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する工程を備えていることを特徴とする請求項１７に記載のテキストマイニング方法。

１９．　テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、

を含み、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項１９に記載のテキストマイニング方法。

２３．（補正後）　テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、

前記文構造の部分構造に対して所定の変換操作を行い、少なくともグラフ構造の枝の繋ぎ変えを含む、前記部分構造と意味的に類似した類似構造を生成する処理と、

前記生成された類似構造を、生成元の部分構造の同値類として扱い、所定のパターン検出を行う処理と、

を実行させるプログラム。

２４．（補正後）　テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、

前記文構造の部分構造に対して、少なくともグラフ構造の枝の繋ぎ変えを含む、予め定められた所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する処理と、

前記生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異を無視しながらパターン検出を行う処理と、

を実行させるプログラム。

２５．　テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、

使用者の入力から、前記文構造の差異の種別ごとに、同一構造と判定するか否かを指定する第１の指定項目と、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成する処理と、

前記文構造の差異の種別ごとに同一構造と判定するか否かを指定する前記第１

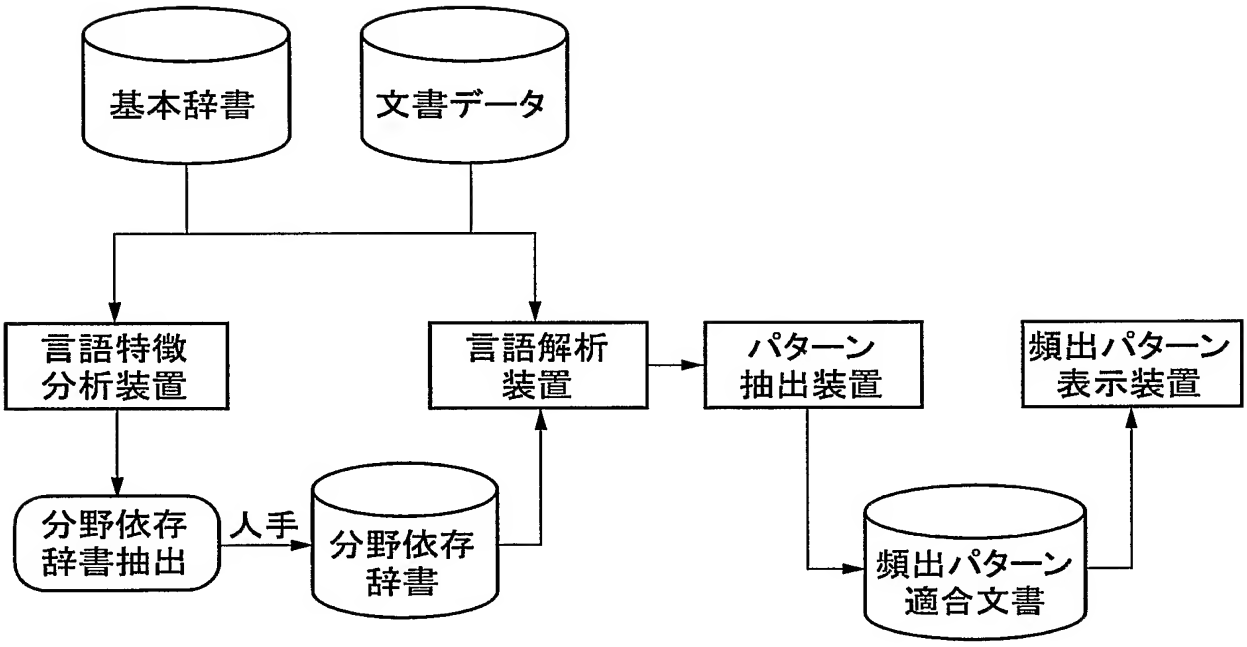

基本辞書
文書データ
言語特徴
分析装置
言語解析
装置
パターン
抽出装置
頻出パターン
表示装置
分野依存
辞書抽出
人手
分野依存
辞書
頻出パターン
適合文書

図 1

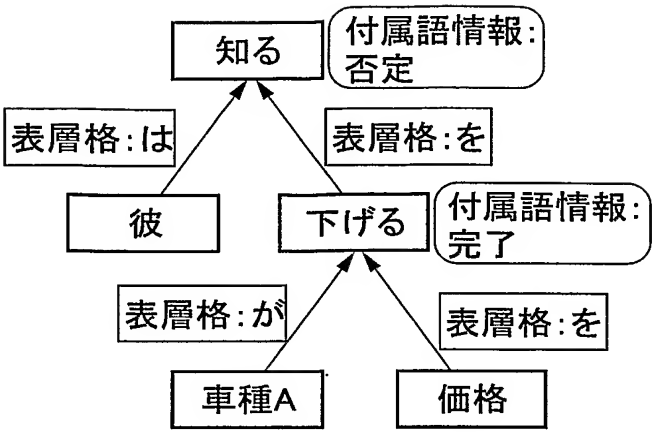

知る
付属語情報:
否定
表層格:は
表層格:を
彼
下げる
付属語情報:
完了
表層格:が
表層格:を
車種A
価格

図 2

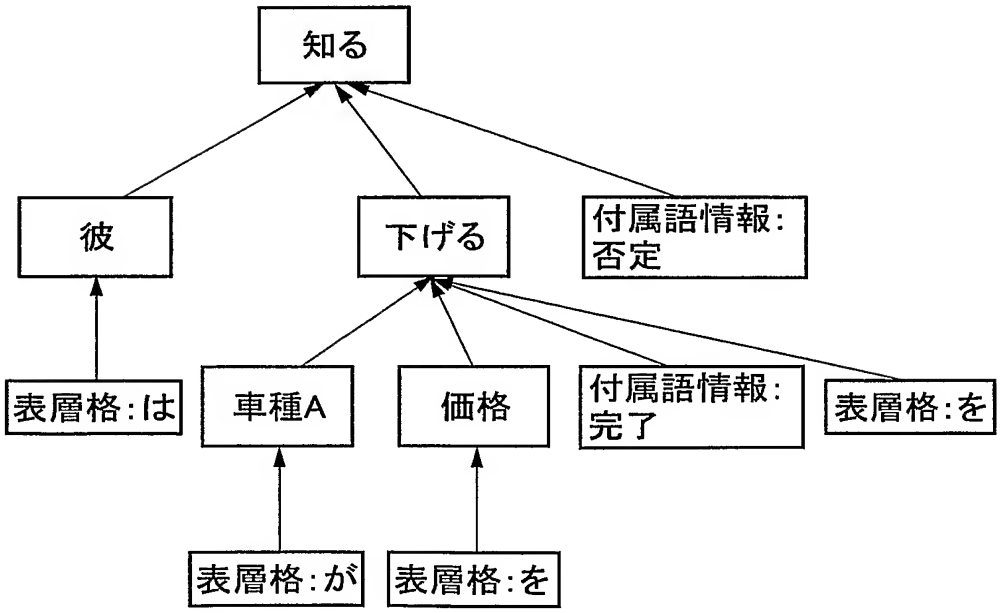

知る
彼
下げる
付属語情報:
否定
表層格:は
車種A
価格
付属語情報:
完了
表層格:を
表層格:が
表層格:を

図 3

「速いのは車種A」と
「車種Aは速い」

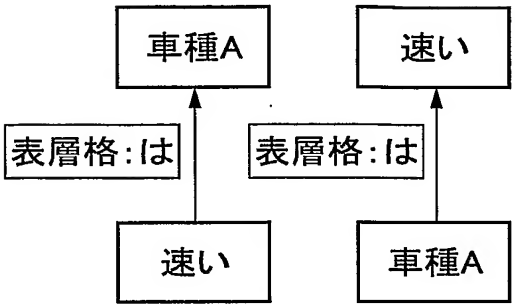


図 4A

「速く安い車種A」と「安く速い車種A」

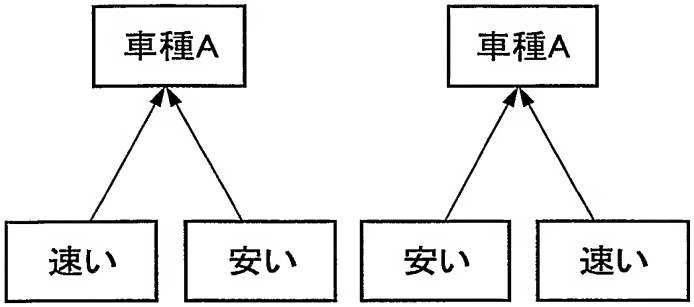


図 4B

「車種Aは加速」と
「車種Aは高速だ」

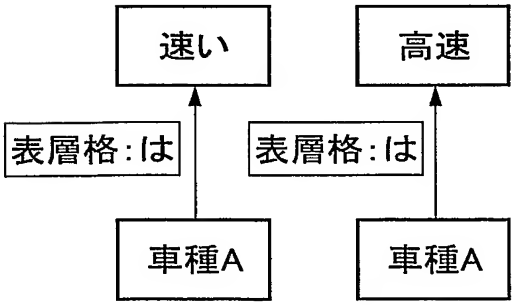


図 4C

「車種Aと車種Bは速い」の
構文構造と意味構造

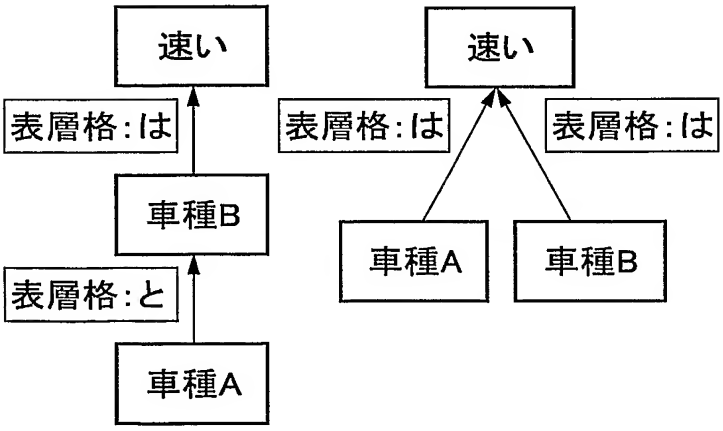


図 4D

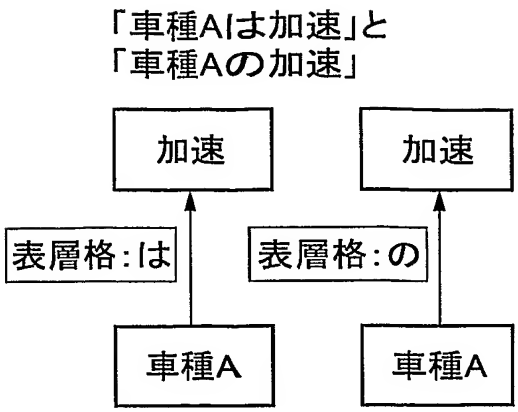

「車種Aは加速」と
「車種Aの加速」
加速
加速
表層格:は
表層格:の
車種A
車種A

図 5A

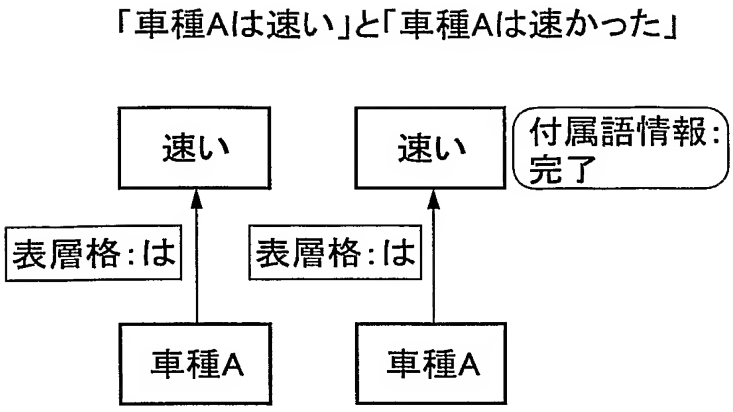

「車種Aは速い」と「車種Aは速かった」
速い
速い
付属語情報:
完了
表層格:は
表層格:は
車種A
車種A

図 5B

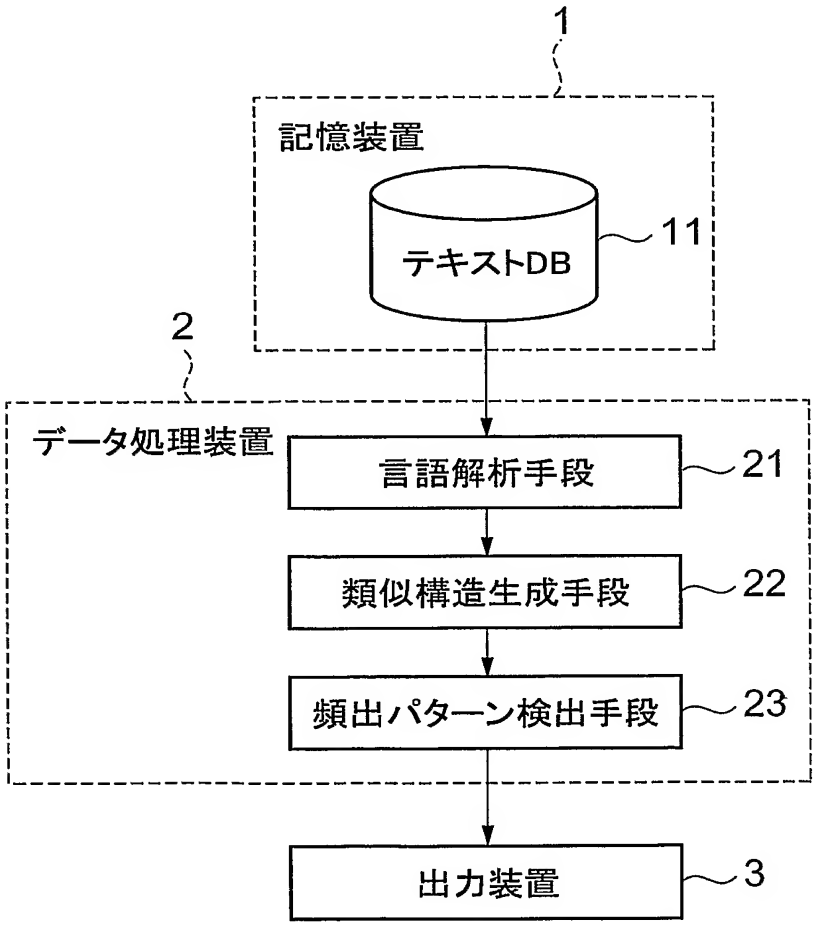

1
記憶装置
テキストDB
11
2
データ処理装置
言語解析手段
21
類似構造生成手段
22
頻出パターン検出手段
23
出力装置
3

図 6

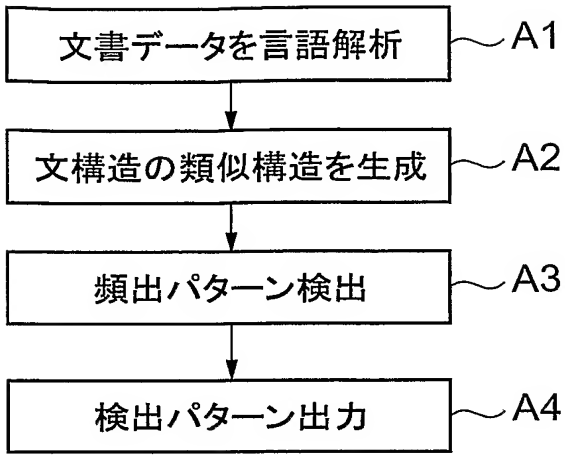

文書データを言語解析
A1
文構造の類似構造を生成
A2
頻出パターン検出
A3
検出パターン出力
A4

図 7

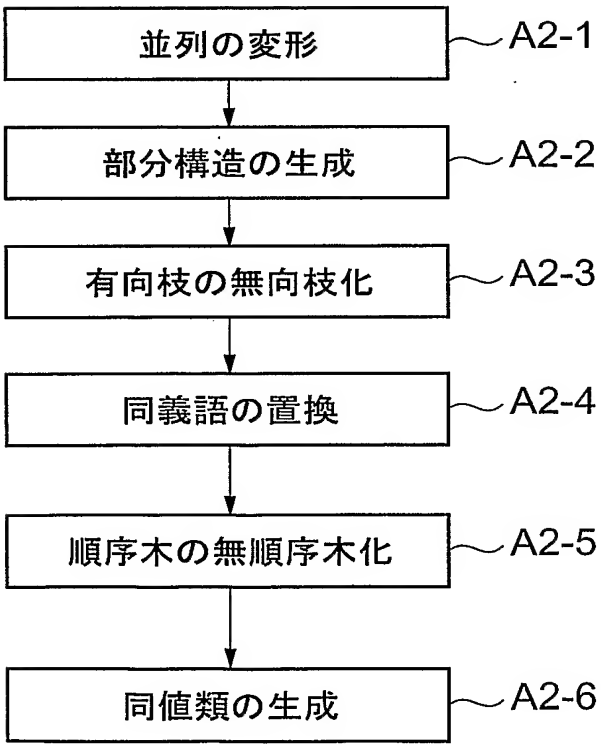

並列の変形
A2-1
部分構造の生成
A2-2
有向枝の無向枝化
A2-3
同義語の置換
A2-4
順序木の無順序木化
A2-5
同値類の生成
A2-6

図 8

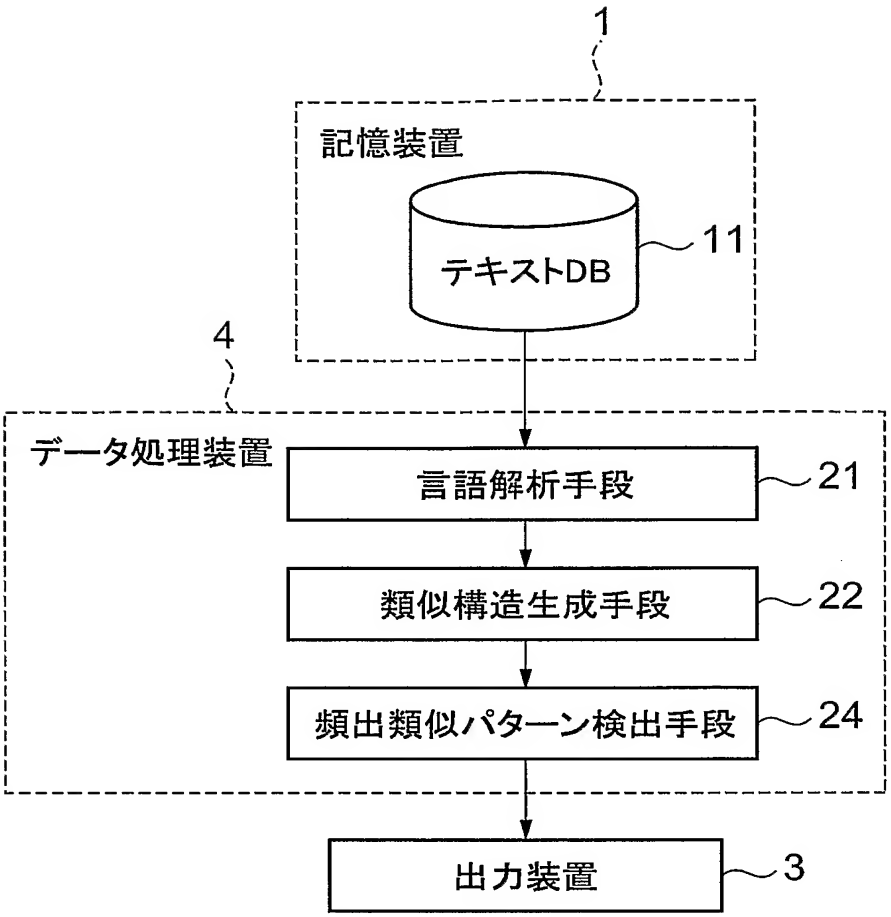

1
記憶装置
テキストDB
11
4
データ処理装置
言語解析手段
21
類似構造生成手段
22
頻出類似パターン検出手段
24
出力装置
3

図 9

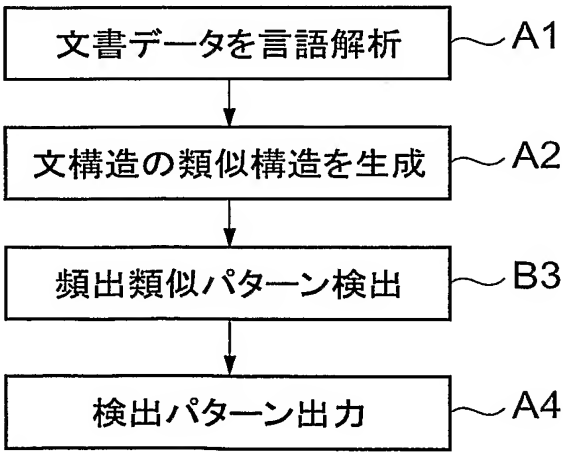

文書データを言語解析
A1
文構造の類似構造を生成
A2
頻出類似パターン検出
B3
検出パターン出力
A4

図 10

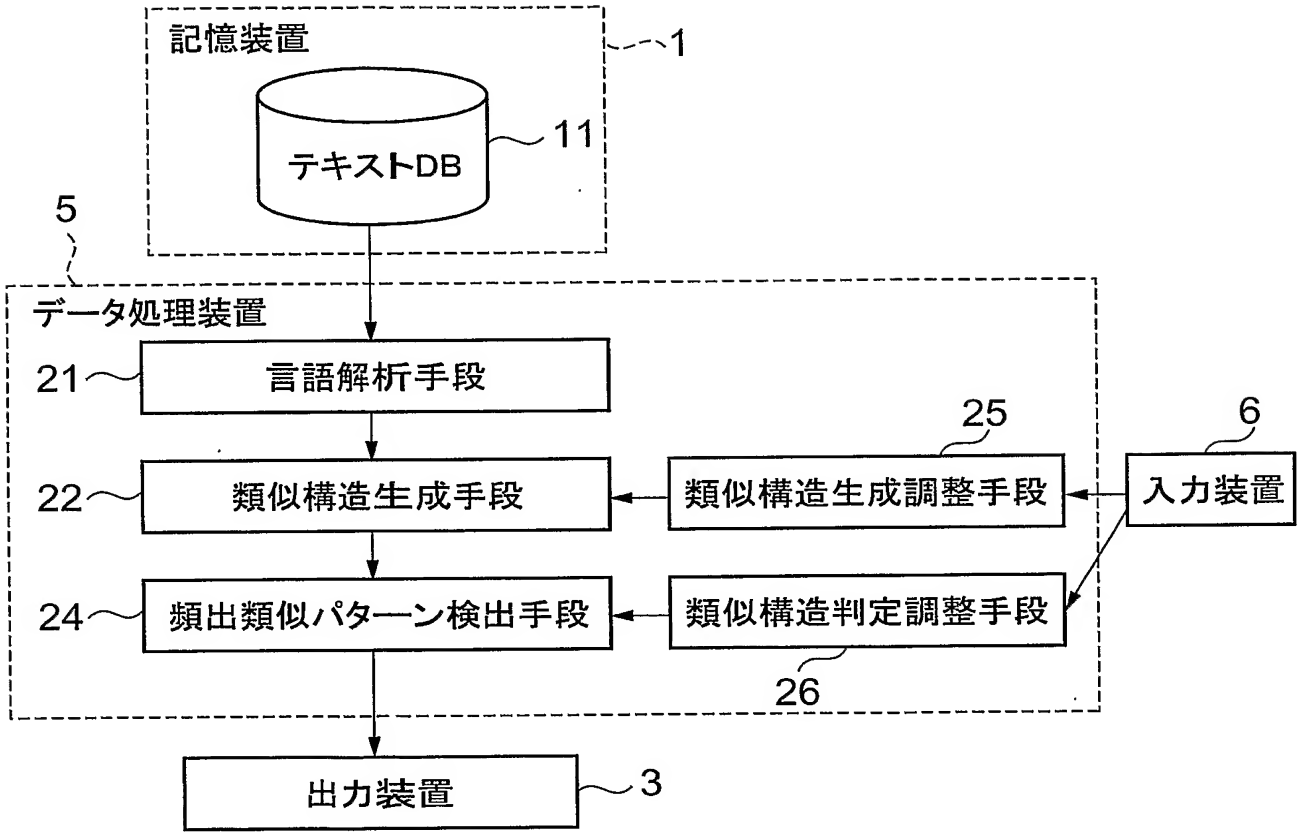

記憶装置
1
テキストDB
11
5
データ処理装置
21
言語解析手段
22
類似構造生成手段
25
類似構造生成調整手段
6
入力装置
24
頻出類似パターン検出手段
類似構造判定調整手段
26
出力装置
3

図 11

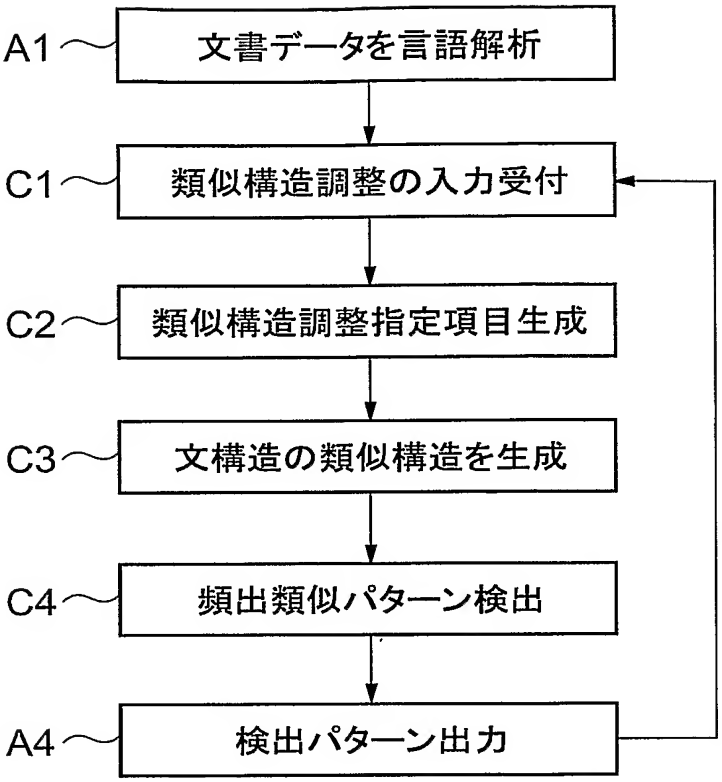

A1
文書データを言語解析
C1
類似構造調整の入力受付
C2
類似構造調整指定項目生成
C3
文構造の類似構造を生成
C4
頻出類似パターン検出
A4
検出パターン出力

図 12

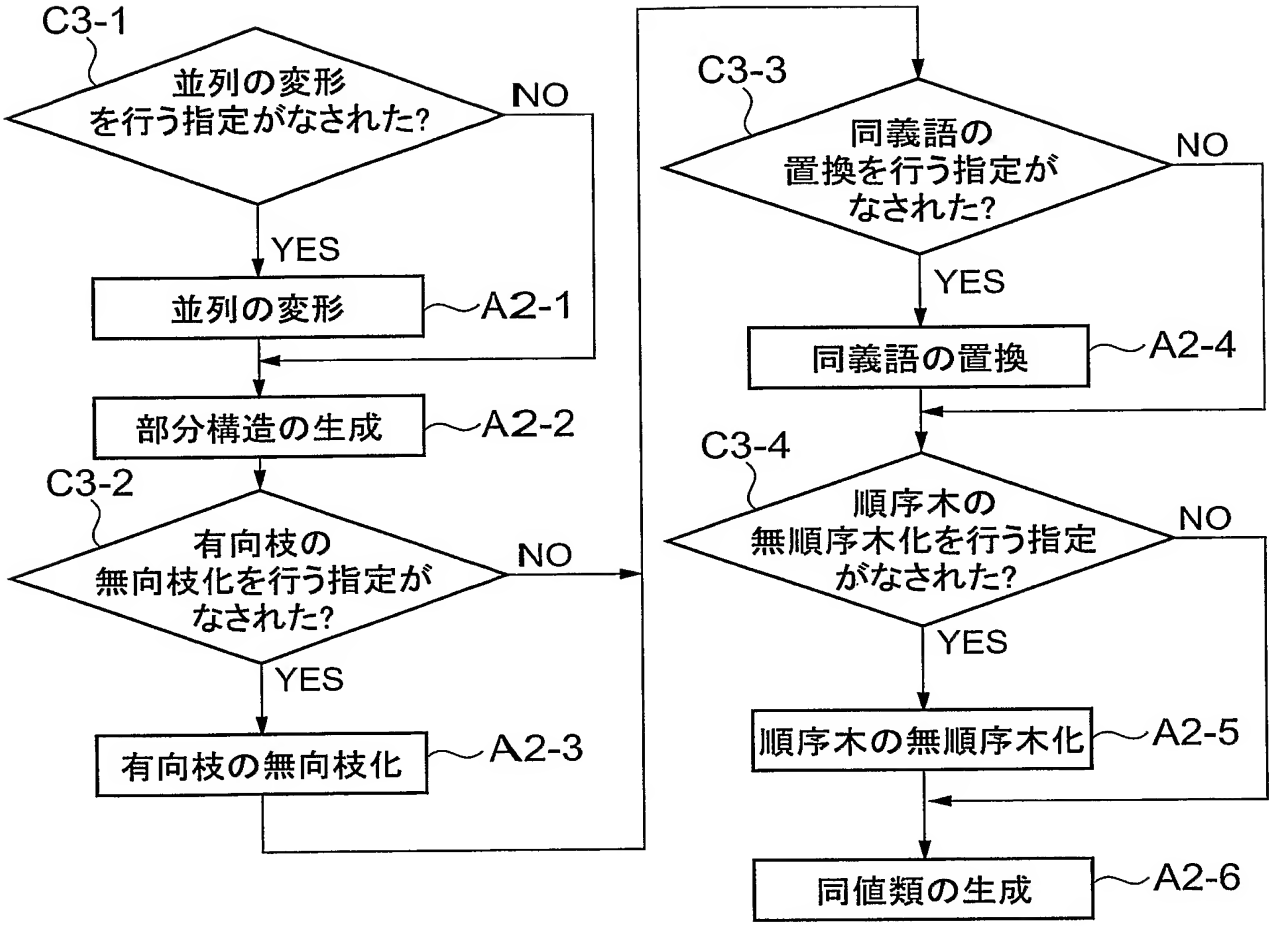

C3-1
並列の変形を行う指定がなされた?
NO
YES
並列の変形
A2-1
部分構造の生成
A2-2
C3-2
有向枝の無向枝化を行う指定がなされた?
NO
YES
有向枝の無向枝化
A2-3
C3-3
同義語の置換を行う指定がなされた?
NO
YES
同義語の置換
A2-4
C3-4
順序木の無順序木化を行う指定がなされた?
NO
YES
順序木の無順序木化
A2-5
同値類の生成
A2-6

図 13

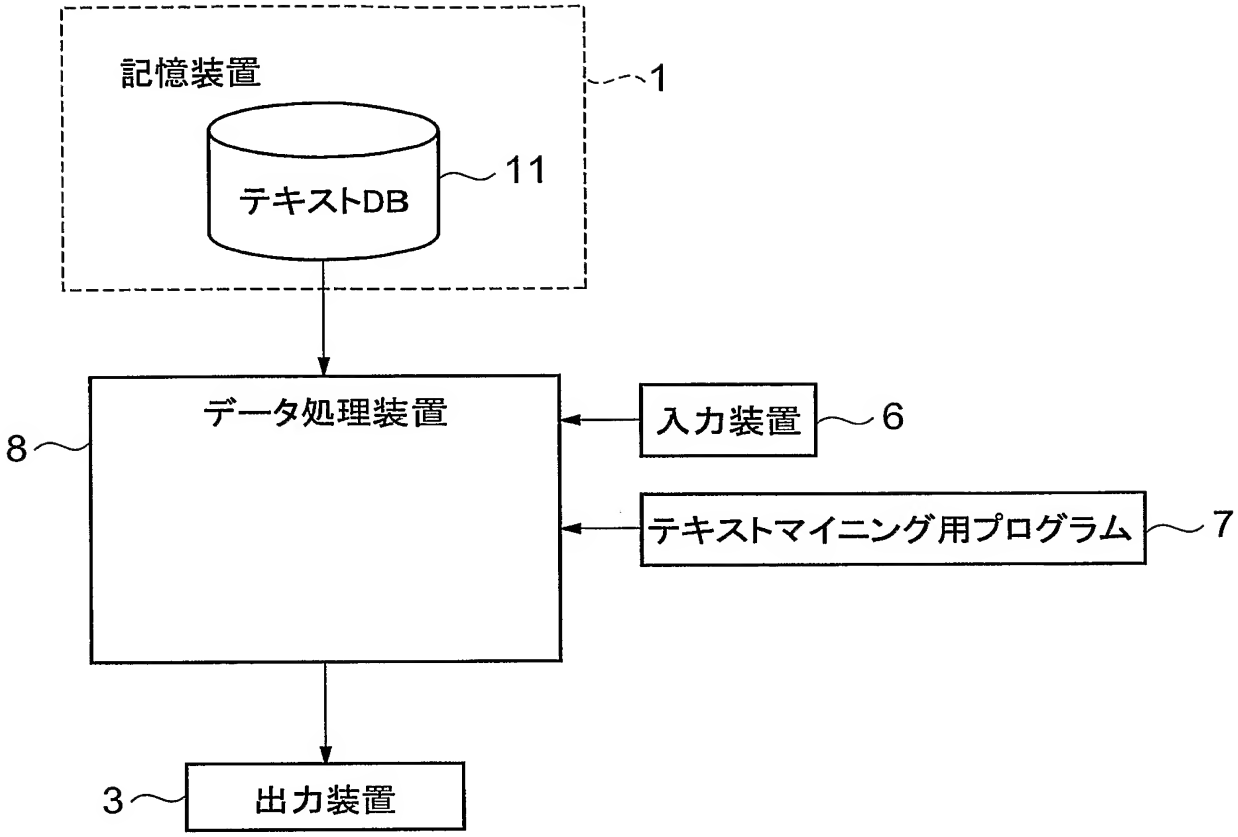


図 14

文1:速い車種Aは安い
文2:速く安い車種A
文3:安かった高速な車種A

図 15

文1の文構造:

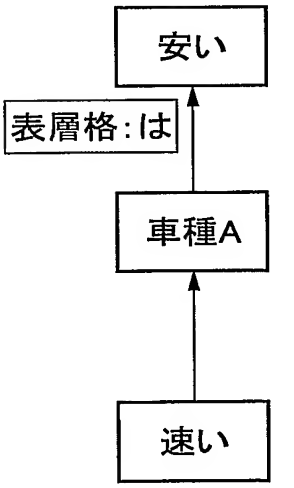


図 16A

文2の文構造:

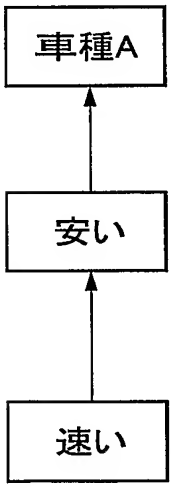


図 16B

文3の文構造:

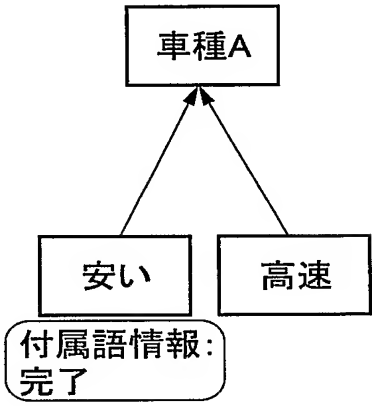


図 16C

同義語辞書

代表語	被置換後
速い	高速

図 17

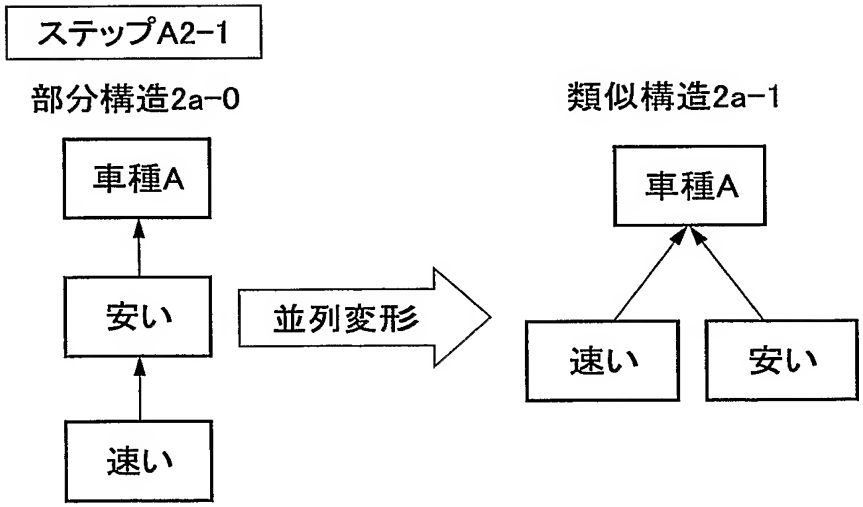


図 18

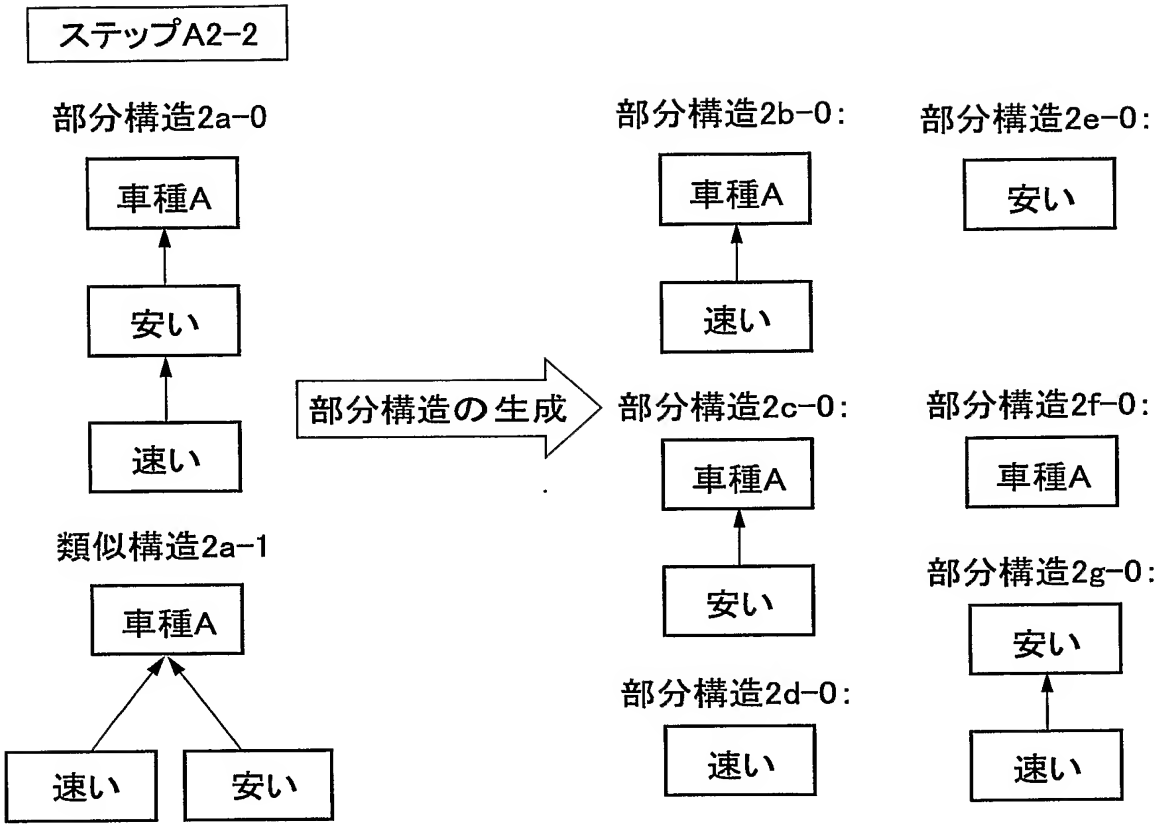

ステップA2-2
部分構造2a-0
車種A
安い
速い
類似構造2a-1
車種A
速い
安い
部分構造の生成
部分構造2b-0:
車種A
速い
部分構造2c-0:
車種A
安い
部分構造2d-0:
速い
部分構造2e-0:
安い
部分構造2f-0:
車種A
部分構造2g-0:
安い
速い

図 19

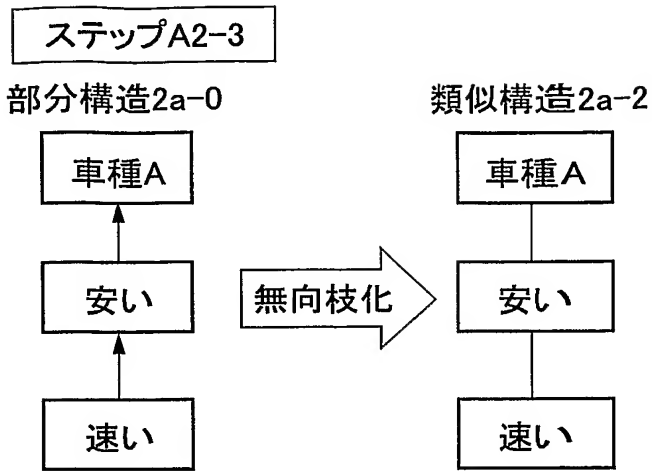


図 20A

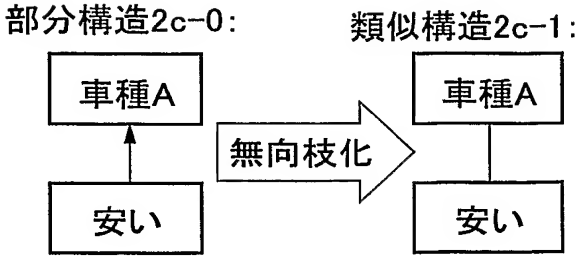


図 20B

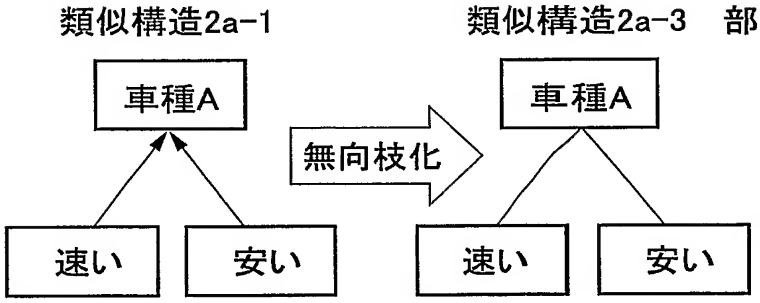


図 20C

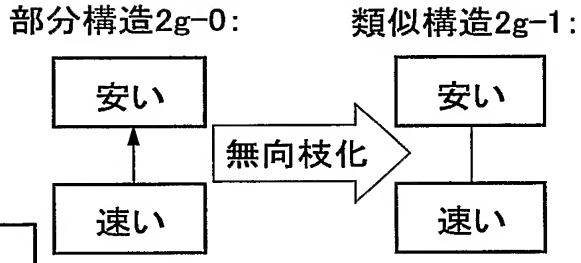
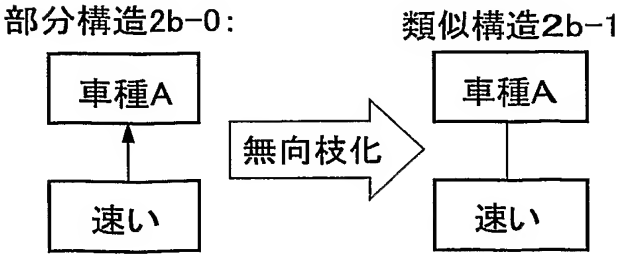


図 20D



部分構造2d-0,2e-0,2f-0はステップ2A-3では形が変わらないので省略している

図 20E

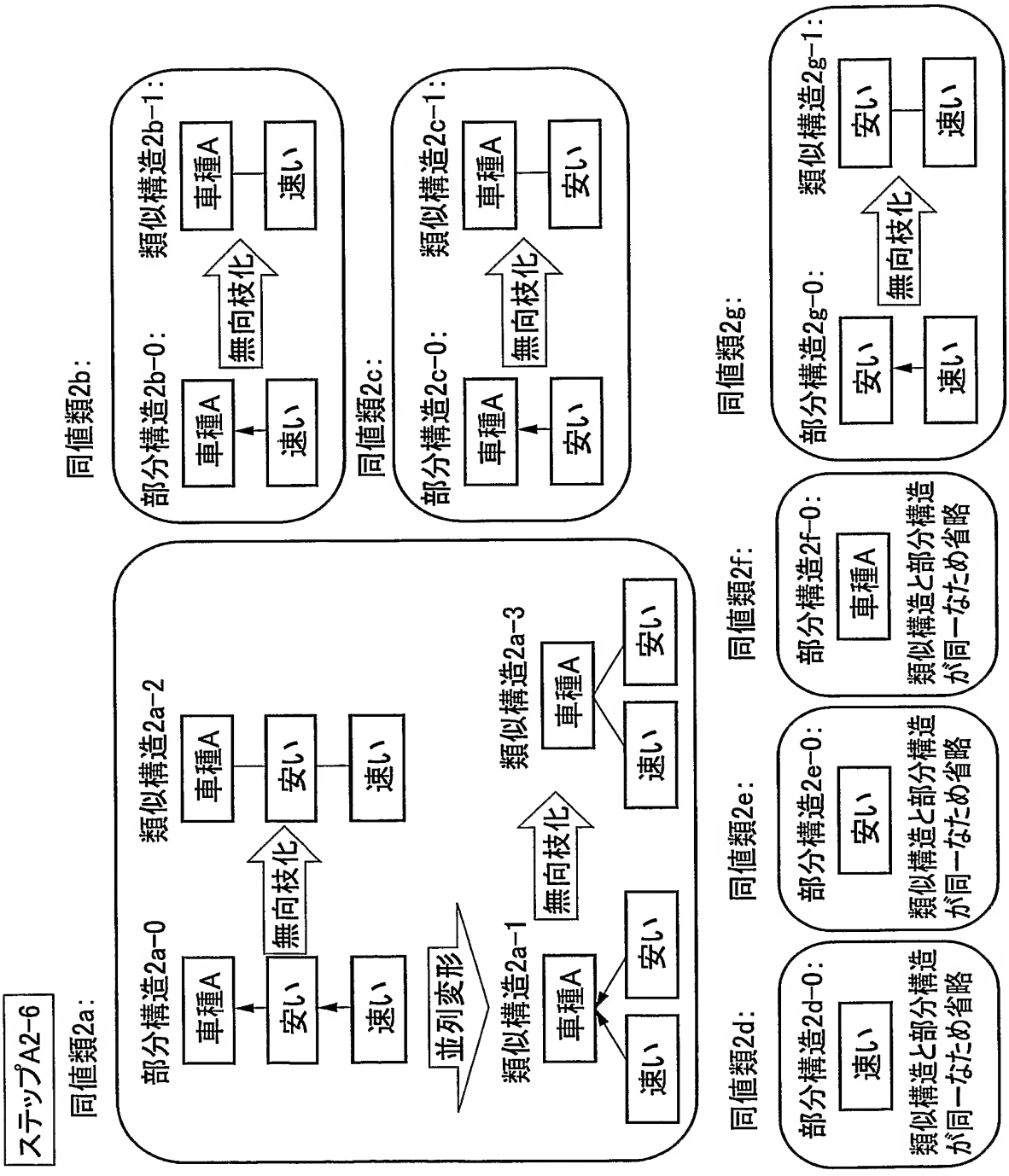

ステップA2-6
同値類2a:
部分構造2a-0
車種A
安い
速い
無向枝化
類似構造2a-2
車種A
安い
速い
並列変形
類似構造2a-1
車種A
速い
安い
無向枝化
類似構造2a-3
車種A
速い
安い
同値類2b:
部分構造2b-0:
車種A
速い
無向枝化
類似構造2b-1:
車種A
速い
同値類2c:
部分構造2c-0:
車種A
安い
無向枝化
類似構造2c-1:
車種A
安い
同値類2d:
部分構造2d-0:
速い
類似構造と部分構造が同一なため省略
同値類2e:
部分構造2e-0:
安い
類似構造と部分構造が同一なため省略
同値類2f:
部分構造2f-0:
車種A
類似構造と部分構造が同一なため省略
同値類2g:
部分構造2g-0:
安い
速い
無向枝化
類似構造2g-1:
安い
速い

図 21

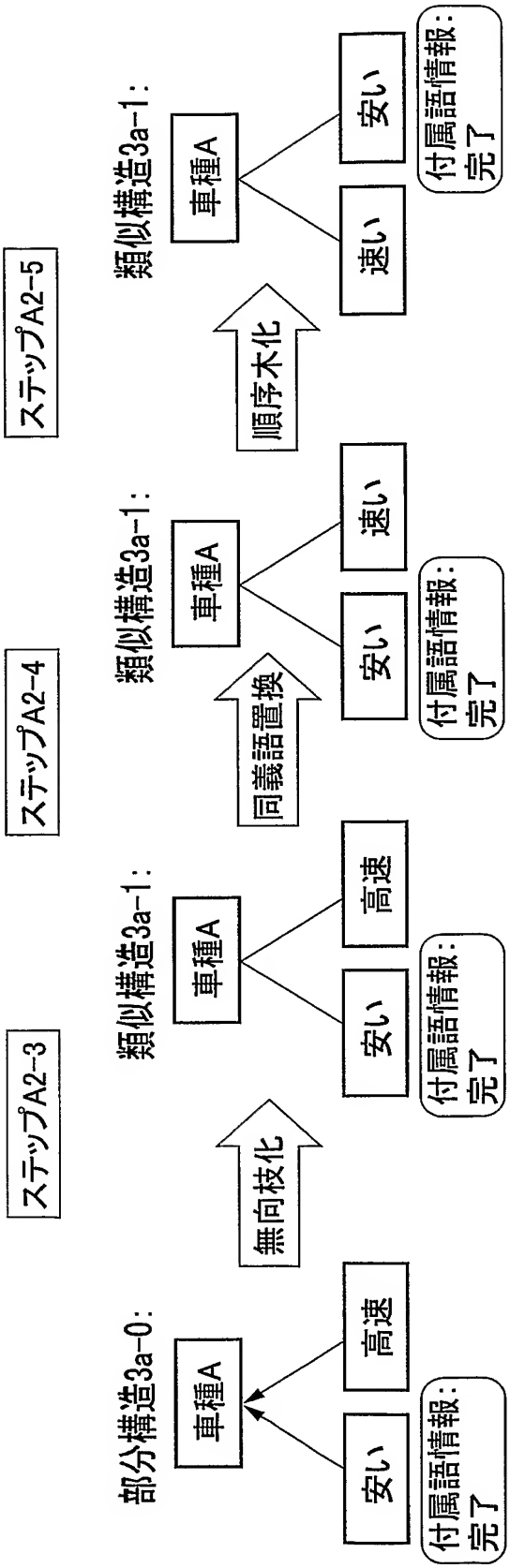

部分構造3a-0:
車種A
安い
高速
付属語情報:
完了
ステップA2-3
無向枝化
類似構造3a-1:
車種A
安い
高速
付属語情報:
完了
ステップA2-4
同義語置換
類似構造3a-1:
車種A
安い
速い
付属語情報:
完了
ステップA2-5
順序木化
類似構造3a-1:
車種A
速い
安い
付属語情報:
完了

図 22

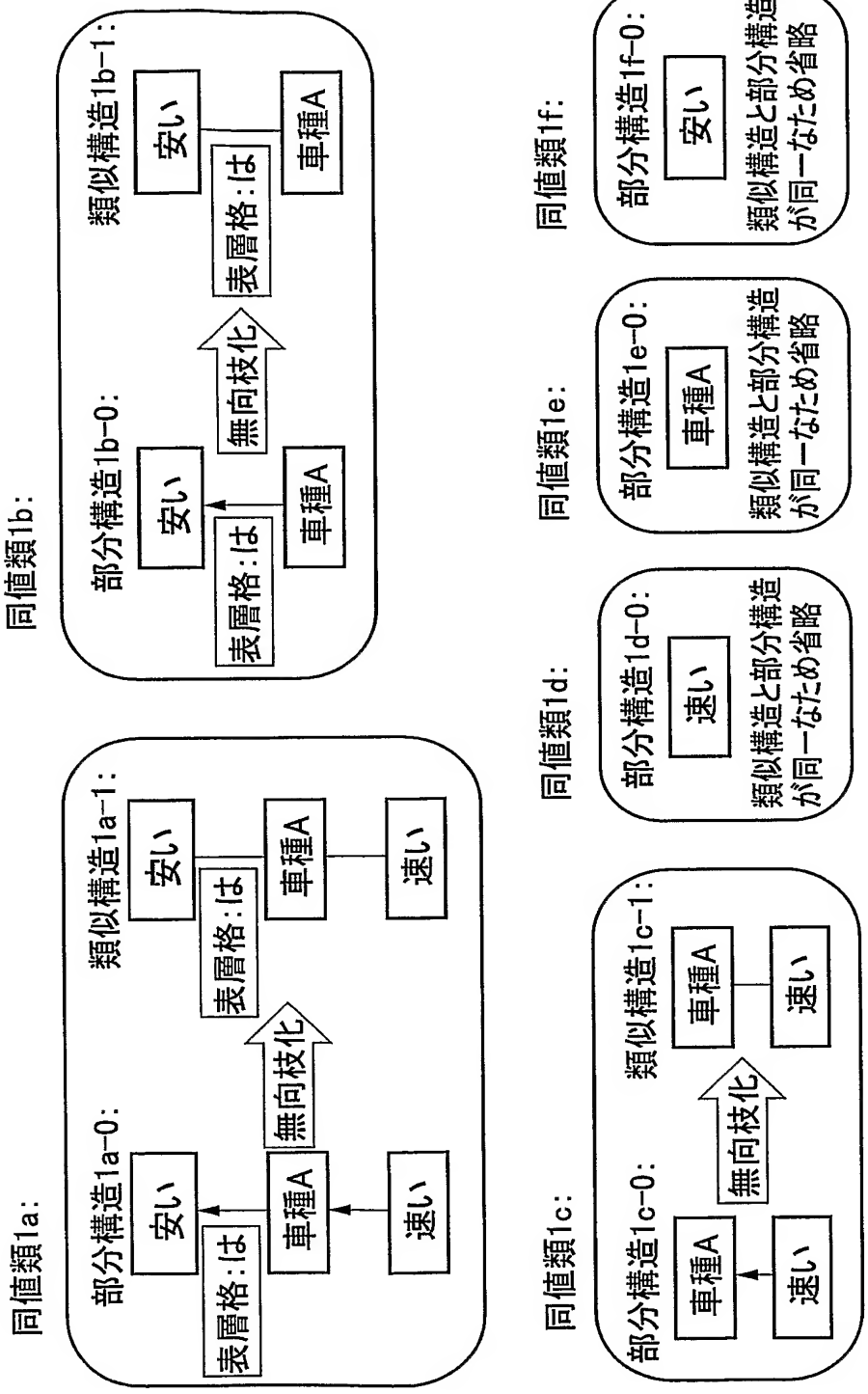

同値類1a:
部分構造1a-0:
安い
表層格:は
車種A
速い
無向枝化
類似構造1a-1:
安い
表層格:は
車種A
速い
同値類1b:
部分構造1b-0:
安い
表層格:は
車種A
無向枝化
類似構造1b-1:
安い
表層格:は
車種A
同値類1c:
部分構造1c-0:
車種A
速い
無向枝化
類似構造1c-1:
車種A
速い
同値類1d:
部分構造1d-0:
速い
類似構造と部分構造が同一なため省略
同値類1e:
部分構造1e-0:
車種A
類似構造と部分構造が同一なため省略
同値類1f:
部分構造1f-0:
安い
類似構造と部分構造が同一なため省略

図 23

同値類2a:
部分構造2a-0
車種A
安い
速い
無向枝化
類似構造2a-2
車種A
安い
速い
並列変形
類似構造2a-1
車種A
速い
安い
無向枝化
類似構造2a-3
車種A
速い
安い
同値類2b:
部分構造2b-0:
車種A
速い
無向枝化
類似構造2b-1:
車種A
速い
同値類2c:
部分構造2c-0:
車種A
安い
無向枝化
類似構造2c-1:
車種A
安い
同値類2d:
部分構造2d-0:
速い
類似構造と部分構造が同一なため省略
同値類2e:
部分構造2e-0:
安い
類似構造と部分構造が同一なため省略
同値類2f:
部分構造2f-0:
車種A
類似構造と部分構造が同一なため省略
同値類2g:
部分構造2g-0:
安い
速い
無向枝化
類似構造2g-1:
安い
速い

図 24

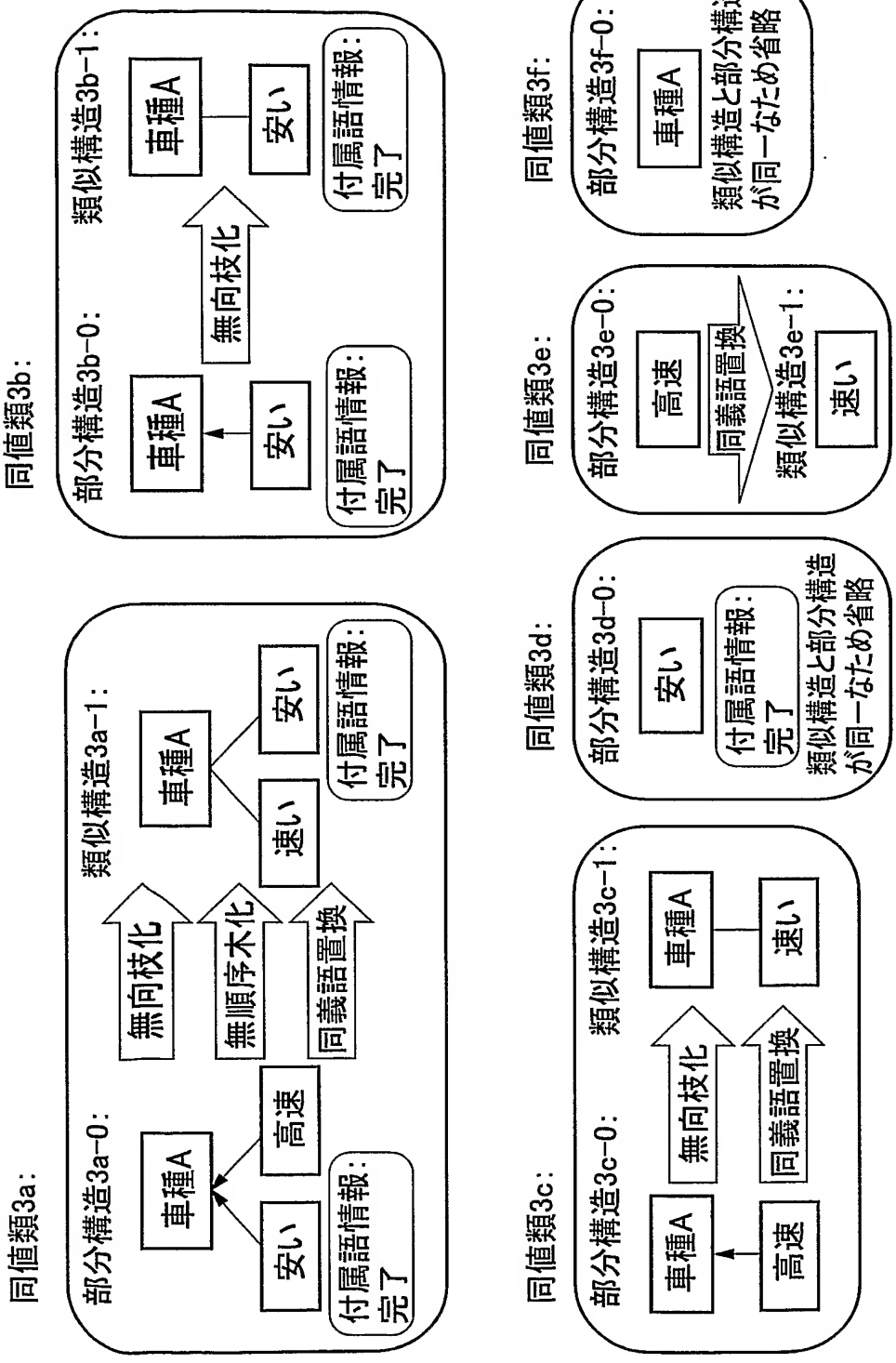
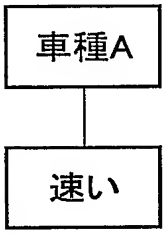

同値類3a:
部分構造3a-0:
車種A
安い
高速
付属語情報:
完了
無向枝化
無順序木化
同義語置換
類似構造3a-1:
車種A
速い
安い
付属語情報:
完了
同値類3b:
部分構造3b-0:
車種A
安い
付属語情報:
完了
無向枝化
類似構造3b-1:
車種A
安い
付属語情報:
完了
同値類3c:
部分構造3c-0:
車種A
高速
無向枝化
同義語置換
類似構造3c-1:
車種A
速い
同値類3d:
部分構造3d-0:
安い
付属語情報:
完了
類似構造と部分構造
が同一なため省略
同値類3e:
部分構造3e-0:
高速
同義語置換
類似構造3e-1:
速い
同値類3f:
部分構造3f-0:
車種A
類似構造と部分構造
が同一なため省略

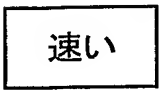
図 25

頻出パターン1:



同値類1c,2b,3cから検出

頻出パターン2:



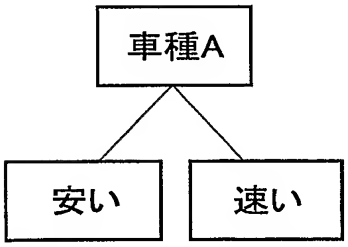
同値類1d,2d,3eから検出

頻出パターン3:

同値類1e,2f,3fから検出

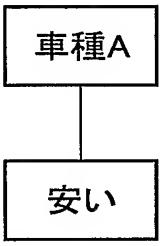
図 26

頻出パターン1:



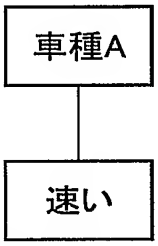
同値類1a,2a,3aから検出

頻出パターン2:



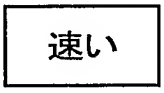
同値類1b,2c,3bから検出

頻出パターン3:



同値類1c,2b,3cから検出

頻出パターン4:

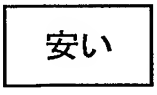


同値類1d,2d,3eから検出

頻出パターン5:

同値類1e,2f,3fから検出

頻出パターン6:



同値類1f,2e,3dから検出

図 27

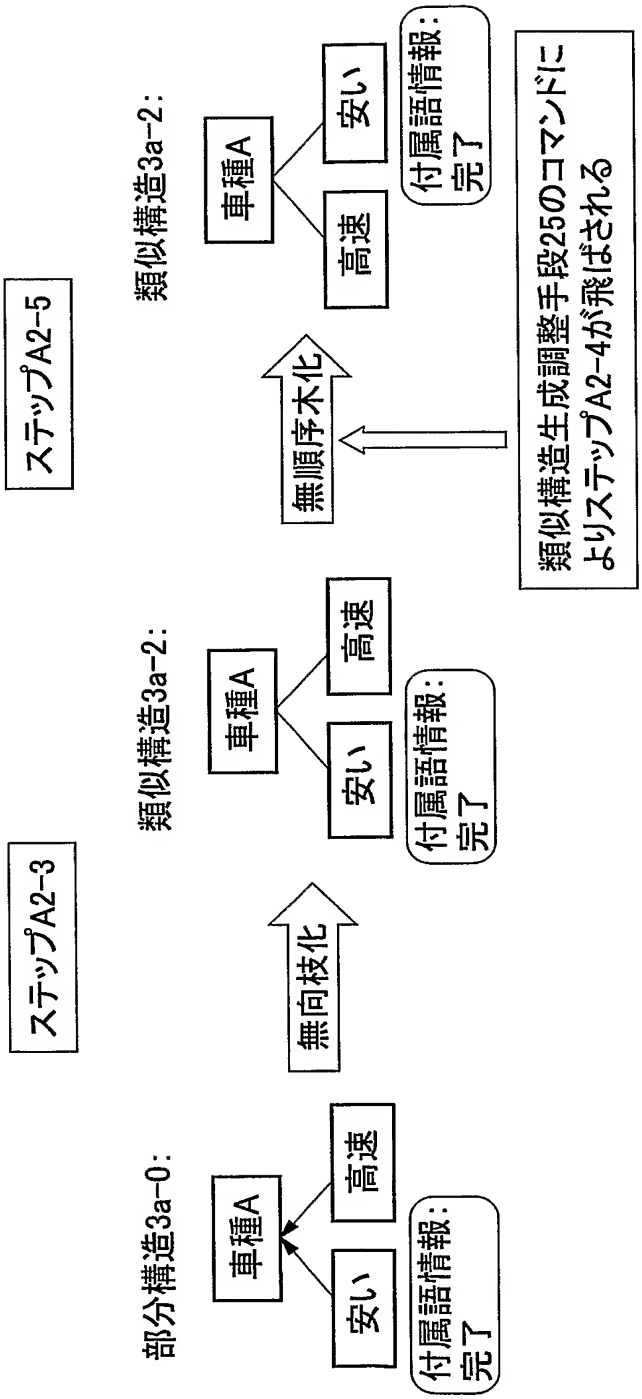

ステップA2-3
部分構造3a-0:
車種A
安い
高速
付属語情報:
完了
無向枝化
ステップA2-5
類似構造3a-2:
車種A
安い
高速
付属語情報:
完了
無順序木化
類似構造3a-2:
車種A
安い
高速
付属語情報:
完了
類似構造生成調整手段25のコマンドにより
ステップA2-4が飛ばされる

図 28

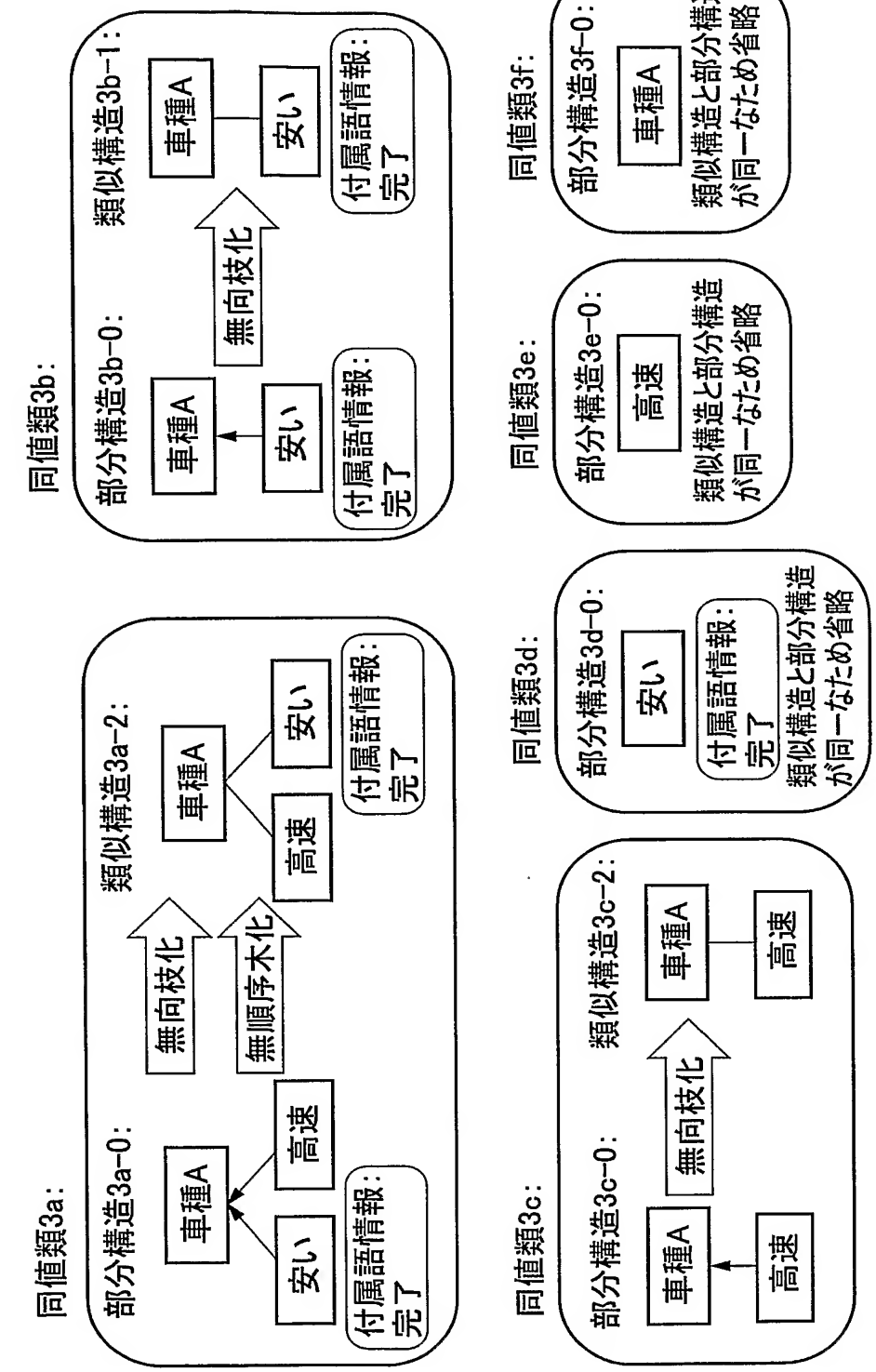

同値類3a:
部分構造3a-0:
車種A
安い
高速
付属語情報:
完了
無順序木化
無向枝化
類似構造3a-2:
車種A
高速
安い
付属語情報:
完了
同値類3b:
部分構造3b-0:
車種A
安い
付属語情報:
完了
無向枝化
類似構造3b-1:
車種A
安い
付属語情報:
完了
同値類3c:
部分構造3c-0:
車種A
高速
無向枝化
類似構造3c-2:
車種A
高速
同値類3d:
部分構造3d-0:
安い
付属語情報:
完了
類似構造と部分構造
が同一なため省略
同値類3e:
部分構造3e-0:
高速
類似構造と部分構造
が同一なため省略
同値類3f:
部分構造3f-0:
車種A
類似構造と部分構造
が同一なため省略

図 29

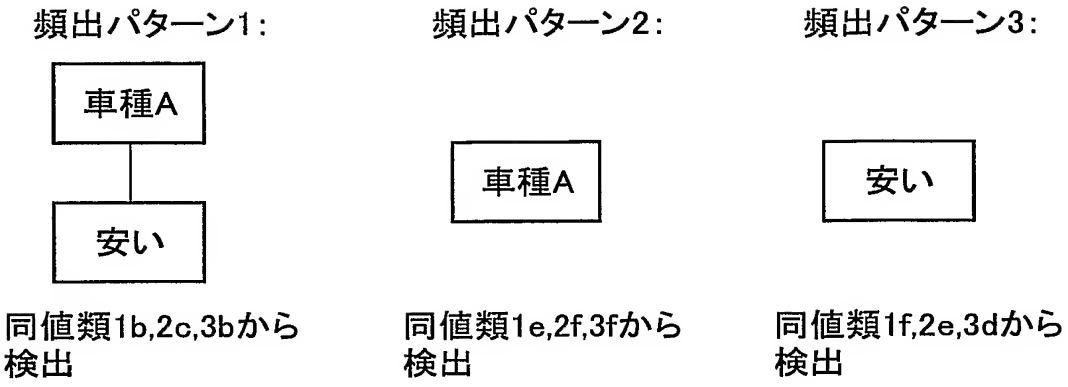

頻出パターン1:
車種A
安い
同値類1b,2c,3bから検出
頻出パターン2:
車種A
同値類1e,2f,3fから検出
頻出パターン3:
安い
同値類1f,2e,3dから検出

図 30

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2005/005440

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/30, 17/27, 19/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/30, 17/27, 19/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2005
Kokai Jitsuyo Shinan Koho	1971-2005	Toroku Jitsuyo Shinan Koho	1994-2005

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus(JOIS)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP 2001-134575 A (International Business Machines Corp.), 18 May, 2001 (18.05.01), Full text; Figs. 1 to 29 & US 6618725 B1	1-7,12-18, 23,24 8-11,19-22, 25
Y A	JP 10-198697 A (Fuji Xerox Co., Ltd.), 31 July, 1998 (31.07.98), Full text; Figs. 1 to 20 (Family: none)	1-7,12-18, 23,24 8-11,19-22, 25
A	JP 2000-76274 A (International Business Machines Corp.), 14 March, 2000 (14.03.00), Full text; Figs. 1 to 4 & US 6219664 B1	1-25

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 April, 2005 (15.04.05)	17 May, 2005 (17.05.05)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2005/005440

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 4-218872 A (Fujitsu Ltd.), 10 August, 1992 (10.08.92), Full text; Figs. 1 to 16 & US 5345537 A	1-25
A	ASAI, T. et al., "Efficient Substructure Discovery from Large Semi-structured Data", Proc. of 2nd SIAM International Conf. on Data Mining, SDM2002, April 2002 [online]; [retrieved on 15 April, 2005 (15.04.05)] Retrieved from the Internet: <URL: http://www.siam.org/meetings/sdm02/proceedings/sdm02-10.pdf>	1-25
A	Taku KUDO et al., "Gengo Joho o Riyo shita Text Mining", Information Processing Society of Japan Kenkyu Hokoku, 2002-NL-148, 05 March, 2002 (05.03.02), Vol.2002, No.20, pages 65 to 72	1-25

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００５／００５４４０

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.[7]　G06F17/30, 17/27, 19/00

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.[7]　G06F17/30, 17/27, 19/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００５年
日本国実用新案登録公報	１９９６－２００５年
日本国登録実用新案公報	１９９４－２００５年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus(JOIS)

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	JP 2001-134575 A（インターナショナル・ビジネス・マシーンズ・コーポレーション）2001.05.18, 全文、第１－２９図 & US 6618725 B1	1-7, 12-18, 23, 24
A		8-11, 19-22, 25
Y	JP 10-198697 A（富士ゼロックス株式会社）1998.07.31, 全文、第1-20図　（ファミリーなし）	1-7, 12-18, 23, 24
A		8-11, 19-22, 25

☑ Ｃ欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日 １５．０４．２００５	国際調査報告の発送日　17.5.2005
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　5M　３２５２ 水野　恵雄 電話番号　０３－３５８１－１１０１　内線　３５９７

様式ＰＣＴ／ＩＳＡ／２１０（第２ページ）（２００４年１月）

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	JP 2000-76274 A（インターナショナル・ビジネス・マシーンズ・コーポレイション）2000.03.14，全文，第1-4図 & US 6219664 B1	1-25
A	JP 4-218872 A（富士通株式会社）1992.08.10，全文、第1-16図 &US 5345537 A	1-25
A	Asai, T. et al., 'Efficient Substructure Discovery from Large Semi-structured Data', Proc. of the 2nd SIAM Internatilnal Conf. on Data Mining, SDM2002, April 2002 [online] ; [retrieved on 15 April 2005] Retrieved from the Internet: <URL: http://www.siam.org/meetings/sdm02/proceedings/sdm02-10.pdf>	1-25
A	工藤拓、外３名，言語情報を利用したテキストマイニング，情報処理学会研究報告 2002-NL-148，2002.03.05，第2002巻，第20号，p.65-72	1-25